从零开始学统计

電子工業出版社
Publishing House of Electronics Industry
北京 • BEIJING

内 容 简 介

大数据时代，每个人都要懂一点统计学，我们缺的不是数据，而是正确分析数据的路径，从海量数据中撷取有用信息、产生新价值，甚至用以推估未知的事物，并且已经成为个人和企业的关键竞争力。这是一本关于统计轻知识的书，作者希望借助轻松幽默的语言来激发读者对统计学的学习热情。内容从描述性统计到推断性统计，通过将生活中有趣的事件一一展开，了解统计学中的核心知识点，最后是常见疑问的答疑汇编。本书偏重于对案例和图表的引用，不会过多关注于数学推导。

本书主要针对未曾学习过统计学或初学统计学并对此有兴趣的读者，以及希望通过学习相关知识补充数据分析技能的在职人士。

图书在版编目（CIP）数据

从零开始学统计 / 归璐编著. —北京：电子工业出版社，2017.1
ISBN 978-7-121-30165-0

Ⅰ. ①从… Ⅱ. ①归… Ⅲ. ①统计学－基本知识 Ⅳ. ①C8

中国版本图书馆 CIP 数据核字（2016）第 254842 号

责任编辑：黄爱萍
印　　刷：三河市华成印务有限公司
装　　订：三河市华成印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编：100036
开　　本：720×1000　1/16　印张：11.5　字数：180 千字
版　　次：2017 年 1 月第 1 版
印　　次：2017 年 1 月第 1 次印刷
定　　价：45.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：（010）51260888-819，faq@phei.com.cn。

学统计的理由

Hi，亲！很高兴遇见你，虽然你看不到我，我也无法目睹你的容颜，但当你翻开这本书的时候，我们就已经通过文字这个载体见面了！

我猜你应该是被本书的标题吸引才会翻开它的吧？那么聪明的你应该知道，这是一本关于统计学的图书。统计学是一门有趣而实用的学科，它将会成为你生活、工作中的好帮手（别告诉我你不炒股、不玩微博、不买彩票，甚至不逛淘宝，你以为我会告诉你这些都和统计有关吗？）。

- 想知道为什么不能赌博吗——学统计吧！
- 想知道为什么淘宝总能“猜透你的心”吗——学统计吧！
- 想知道怎样才能获得升职加薪的捷径吗——学统计吧！

你有没有想过买一张福利彩票，然后被五百万元大奖砸中？我就想过，那通常发生在大白天，两眼呆滞且目光涣散，幻想自己抱着一堆红色的人民币傻乐……但是当我回过神来之后，我就清楚地意识到中大奖的机会微乎其微——这是概率论教会我的。

你也许会想：这是我小时候就懂的道理，你还要读了概率论才知道？

要知道，概率论诞生于赌博游戏。一两次的小赢，甚至接连几次

都赢是有可能发生的，这属于概率的正常波动。其实，如果在完全公平的情况下，输赢概率应该各为 50%。但为什么总感觉赌的时间越长，越容易输呢？这是因为我们忽视了一个重要的因素，那就是输赢各半的前提是可以进行**无限多次**的赌博，但事实上我们根本不可能有那么多的资金和精力。要知道，得出抛硬币正反面出现概率各为 50%的结论，是建立在上万次试验结果之上的。所以，你若知道**概率还蕴含积分**的数学思想，就不难理解为何“十赌九输”了。

你有没有想过，“万能”的淘宝为何总能在你搜索宝贝的时候顺便推送一些名为“猜你喜欢”的产品，而且这些推送有时还能被你成功加入购物车？其中就用到了推荐算法。推荐算法不仅涉及文本挖掘技术，而且与统计学中频率的计算和关联性知识有紧密联系。

在我们的日常工作中，如果你从事的是销售、财务工作，或者你是某项目的策划者，当领导询问你对即将上架的产品，或者要削减某项开支，或者某项目的推广方案的看法时，你该如何回答？

如果你对自己所做的工作有过翔实的数据采集，例如，对需要销售的产品做过统计，就可以得出一系列图表来证明该产品在某个时间段或针对某些特殊人群有明显的销量提升（这通常涉及方差分析）；再如，你对公司的财务数据做了详细的台账记录，则可以清楚地知道缩减哪些开支既不影响生产销售又可以提高营业利润（这时可以运用相关分析）；又如，你使用定量方法将推广方案的定性数据量化，通过分析得出最佳方案。试着使用数据来说话，慢慢培养统计思维，你会发现，你的工作将会事半功倍。

生命和统计息息相关

如果上述例子无法给你学习统计的充分理由，那么，当数据和生命联系在一起时，会是怎样的呢？

手术中，麻醉师的用药剂量与病人的个体情况有着严格的匹配要求；新药物上市之前，必须经过无数次试验检验；用药说明书上的剂量指导，更是建立在海量试验检验基础之上的。其中就涉及抽样调查、假设检验和实验设计等多种统计学的理论知识。

不久前，“雾霾致癌吗”这个话题异常火爆。关于这个命题的真伪，在此不做评述，但众所周知，吸烟是有害健康的，吸烟致癌也被大家广为接受。但你知不知道，“吸烟是否是引起肺癌的原因”这个论题曾经在统计学界掀起了轩然大波？当时，费希尔（统计学界的泰斗级人物）极力反对这个观点，其实，在证明吸烟与肺癌关系的过程中，更值得讨论的是对于试验的设计和流行病医学里的因果关系的论证。直到目前，仍然没有一种有效的方法能够证明统计学和哲学双层面的因果关系。但随着统计学的飞速发展，医学统计逐渐流行起来，并发展成为一门热门学科。

生活中的每一部分都和统计密切相关

当一门学科发展到可以通过量化数据来解密人体科学的时候，还能说它不值得去学习了解吗？比如，在大数据时代，如果你不会两个统计名词，怎能充分利用大数据的价值？从事金融行业的不会数据分析，不能跑代码，怎么体现你的专业素养？如果没听说过什么是Hadoop/R/SAS，你怎么做合格的程序员？还有机器学习、词频分析、文本挖掘、数据挖掘……所有这些都离不开统计理论的支撑。所以，

如果你想走在时代的前沿，就抓紧时间学统计吧！

当然，即使有千万个学习统计的理由，但总有一个理由会让你拒绝学习，那就是数学！你不热爱数学，所以拒绝学习和数字有关的学科。但是，这并不能成为你不学习统计的理由，因为统计和数学并不相同。我认为，统计学就是“高冷”数学和深奥哲学的平衡点。

其实，我天生对数学也没有兴趣，丝毫看不出那些积分符号优美在何处。但是这并不能阻碍我对统计学的热爱。诚然，统计理论是完全建立在数学基础上的，数理统计对数学的要求很高，但是统计学里还有一个分支叫应用统计，本书就是为了应用而生的。

本书不会有繁冗的数学公式推导，不过在有些时候，为了说清楚问题，数学公式和定理是不可或缺的。水平有限，力争通过通俗易懂的语言让大家明白统计是怎么回事，以及统计可以用来做些什么。

你不用惧怕巨大的计算量，这些都可以通过软件来完成。喜欢编程并想深入研究理论知识的，可以使用 Stata、SAS、R；想要快速解决问题的，可以使用 SPSS；甚至可以使用 Excel 完成绝大多数统计分析工作。

至此，你应该找不到不学统计的理由了吧？

欢迎大家和我一起进入奇妙的统计学世界！

归　璐

2016 年 12 月 1 日

目　录

第 0 章　入门阶段——带你迈入统计学的大门 …… 1

0.1　我和统计学的从零开始 …… 1
0.2　统计学的从零开始 …… 4

第 1 章　你的数据从何而来 …… 12

1.1　“不可能完成的任务”——普查 …… 13
1.2　“四两拨千斤”——事半功倍的抽样调查 …… 15
☆本章重点归纳 …… 22

第 2 章　掌握指标学会数据分析 …… 24

2.1　被误解还是“被平均” …… 24
2.1.1　数值平均数——最熟悉的陌生人 …… 26
2.1.2　位置平均数——关键的排序 …… 31
2.2　均值的好朋友——方差（标准差） …… 37
2.3　峰度&偏度——打造风度翩翩的数据分布 …… 41
☆本章重点归纳 …… 44

第 3 章　图表的世界 …… 45

必备技能 1——频数分布表 …… 45
必备技能 2——频数分布图 …… 49
必备技能 3——茎叶图 …… 52

必备技能 4——箱线图 ……55
必备技能 5——散点图 ……58
☆本章重点归纳……65

第 4 章　当小“正太”遇上“大叔”——正态分布篇……67

4.1　小“正太”的基本情况 ……68
4.2　小“正太”的性格和优点——正态分布的定义和特征 ……69
4.3　小“正太”的可爱之处——正态分布的作用……72
☆本章知识点补充……79

第 5 章　当小“正太”遇上“大叔”——大数定律和中心极限篇……81

5.1　正态分布的“左膀”——大数定律……81
5.2　正态分布的“右臂”——中心极限定理 ……84
5.3　如何牵手“大叔”和“正太” ……88
☆本章重点归纳……89

第 6 章　相关和因果切莫傻傻分不清楚……91

6.1　为了“不确定”的确定 ……92
6.1.1　散点图 ……93
6.1.2　相关系数 ……95
6.2　上帝掷骰子 ……101
☆本章知识拓展……103

第 7 章　“小”亦可为，“大”而佐之 ……106

7.1　这个“小二”一点都不“二” ……106
7.2　另辟蹊径的最大似然估计法……110

7.3 他山之石，或可攻玉 …… 113
☆本章知识拓展 …… 115

第 8 章 从先放牛奶 or 先放热茶说起 …… 117

8.1 掀开假设检验的面纱 …… 119
8.1.1 原假设 VS 备择假设 …… 120
8.1.2 检验统计量和拒绝域 …… 123
8.1.3 *P* 值 …… 126
8.2 几种常用假设检验简介 …… 128
8.3 手把手教你做检验 …… 131
☆本章知识拓展 …… 135

第 9 章 回归分析——科学研究的“万金油” …… 137

9.1 探寻“回归”的本质 …… 138
9.2 释放“回归”的超能力 …… 141
9.3 规避“回归”的误区（伪回归问题） …… 146
☆本章知识拓展 …… 149

第 10 章 物以类聚，人以群分 …… 152

10.1 分久必合——聚类分析 …… 152
10.2 合久必分——判别分析 …… 158

第 11 章 独辟蹊径，曲径通幽 …… 163

第 0 章

入门阶段——带你迈入统计学的大门

0.1 我和统计学的从零开始

既然书名是《从零开始学统计》，那么本书的目录自然也从第 0 章开始。0 意味着起点，在我们开始系统地了解统计学之前，先来听我讲讲我和统计学之间的故事。

我和统计学的相识是一场美丽的意外。在选择统计学专业之前，我对统计的了解仅限于求平均数、求方差。如果说得再深奥一点，那么还能略微扯上一些概率论。对于学了统计学将来能做什么，我也是一知半解。是什么原因让我选择了这个在当时略显生僻的专业呢？原因很简单——好奇。

“统计”一词起源于国情调查，最早意为国情学。首先来看看“统”字的含义。“统”字可以作三种解释：（1）充满、充盈；（2）总括，

总起来，如统一、统帅等；（3）事物的连续关系，如系统、传统等。从中可以看出，统计学的“统”更倾向于后两种解释。“计”为核算之意。那么两者相结合，表示对总体的核算和对事物连续关系的计量。结合日常生活，一些工作偏向于总体的核算，如对宏观经济数据的披露；而现如今一些职业如 Data Scientist 则需要统计学的专业背景，且更倾向于事物连续关系的挖掘。两者有一定的共性，归结起来就是统计的定义：对数据进行收集和整理，并在此基础上加以分析和科学决策。至于怎么收集和整理数据、怎么分析和决策，将在本书的后续章节详细介绍。

客观地说，数学功底好对于学习统计学大有益处，但这并不能保证你一定能够学好统计学。以笔者的经验来看，**统计学真正迷人的地方在于统计方法和统计思想**。在很多优秀的统计学著作里，通常看不到长篇大论的数学证明，有些甚至放在附录中，正文则更多地阐述数据处理方法的创新，以及建模和算法的创新。

为什么说数学好未必能学好统计学呢？首先，**数学讲究严密的逻辑演绎，而统计学则更多的是归纳推理。**比如，通常人们认为，统计结论都应该建立在数据服从正态分布的基础之上，但很多数据仅仅是近似服从。这么宽泛的条件，怎么能得到让人信服的结论？笔者试图用大数定律和中心极限定理来验证结论的可信度，但事与愿违。其中的矛盾就在于统计学往往更注重应用。在实际应用中，数据是无法达到完美的理论要求的，适当地放宽和采用近似方法反而更能够接近真相。

其次，市面上**种类繁多的统计软件**，让那些不擅长数学的人也

可以掌握统计学的知识。常用的统计软件有：龙头老大——SAS；后起之秀——R；新手福音——SPSS；擅长面板数据计量分析的 Stata/MATLAB；计量入门小能手 Eviews；数据挖掘方面也有 Clementine、Python 等。

如果你不想深入研究，只想利用统计学来解决一些非统计专业领域的难题，那么，大可不必选择高深的软件，拥有菜单操作的 SPSS 甚至 Excel 都可以满足你的统计需求。是的，只需轻轻地单击一下，结果自然呈现。但前提是你必须知道结果的含义，也知道如何选择正确的统计方法。

但如果你想要专业一些，那么还是需要学习 R、SAS 和 Python 的。R、SAS、Python 是目前比较热门的软件，通常金融类企业需要处理海量数据，SAS 使用频繁，而且较为权威；R 是免费开源的，包含各类程序包，所以现在很多分析公司也会采用 R 作为主要软件，也有很多编程爱好者喜欢研究 R，如果你的工作偏向于数据分析类，那么 SAS 和 R 可以任取其一；如果你的工作偏向于数据挖掘方向，那么可以考虑选择 Python，它的应用面非常广。

学习统计软件的过程不仅仅是为了简化运算，也不单单是为了建模。笔者之所以喜欢统计，很大一部分原因在于在学习这些软件的同时加深了对统计思想的理解。笔者通常会把数据在各类统计软件里执行一遍，看结果会有何不同；也会试着用不同的检验方法检验同样的数据，如使用参数检验和非参数检验，再来对比一下结果有何不同。尤其是在进行多元统计分析的时候，如进行聚类分析，不同的数据处理方法会带来完全不同的结果。这类小实验给笔者的统计学习带来很

大的乐趣。

统计学是一门探索的学科，一百个人做同一个统计研究可能得出一百个结论。但同样的，统计学也带给你更多的提问机会。学好统计学并不难，只要你喜欢问为什么，也喜欢去回答为什么就可以了。

笔者认为统计学有着“中庸”之美！“中庸”并非数学中的中项，恰恰是精确可计的两端的平均数，它随着环境的改变而改变，并且只对成熟且有灵活性的理性才显露自身。精度与费用之间的平衡就是“中庸”的体现。要知道，误差是统计学的一个特征，如果你不能跳出这个思维限制，过度纠结于理论的严苛条件，那么就很难学好统计学了。

统计学还有着“哲学”之美！它是一种由经验到理性的认识，是一种运用偶然性来发现规律性的科学。偶然中蕴含着必然，这属于统计学的哲学美，这个美的最佳体现就是大数定律。

0.2 统计学的从零开始

比起古老的数学（初等数学诞生于公元前 5 世纪），统计学也可以算作一门有着浓厚历史文化的学科，追溯其源可以发现，统计学和亚里士多德有着千丝万缕的联系。历史悠久的统计学经历了人类的农业经济时代、工业经济时代，并在知识爆炸的今天掀起了一次新的“生长发育”。

要说统计学的发展史，不得不说它名字的由来。“统计学”一词最早来源于现代拉丁文 statisticum collegium（国会）。那时，亚里士

多德写了 150 多种纪要，这些纪要被称为“城邦纪要”，其内容包括各城邦的历史、行政、科学、艺术、人口、资源和财富等社会和经济情况的比较分析，具有社会科学的特点。到了 16 世纪，意大利语用 statista 来称呼和政府相关的政治家；接着，德国人戈特弗里德·阿亨瓦尔开始使用 statistik 一词来表示对国家资料进行分析的学问；1785 年，在法语中出现“统计”一词，写为 statistique；1807 年，丹麦语也引入 statistik 作为统计的名称；最终演化为现如今的“统计学”（Statistics），依然保留了城邦（state）这个词根。

任何一门学科在其发展的道路上都会有派别的划分和争斗，统计学也未能幸免。在其发展道路上，每一次衍生出的新派别都是推动学科前进的动力。

1．17 世纪——政治算术学派 VS 国势学派

（1）政治算术学派——统计学的始祖：威廉·配第&约翰·格朗特。

17 世纪，在英国诞生了政治算术。这里的“政治”是指政治经济学，“算术”是指统计方法。其代表人物之一是威廉·配第，如图 0.1 所示。

威廉·配第出生于英国的一个手工业者家庭，早年学过数学、希腊文和拉丁文，接着去法国继续深造数学、天文和航海，后在皇家海军中服役，又到巴黎和阿姆斯特丹学习医学。他的后半生是在爱尔兰度过的，在那里，他主持土地丈量的工作，并与爱尔兰的一些政治和经济问题有过关联。晚年成为拥有大片土地的大地主，还先后创办了

渔场、冶铁和铝矿企业。威廉·配第在其代表作《政治算术》一书中写道："本书不用比较级、最高级进行思辨或议论，而是用数字来表达自己想说的问题，借以考察在自然中有可见的根据的原因。"该书标志着统计学的诞生。

图 0.1　威廉·配第

在这本书中，威廉·配第利用实际资料，运用数字、重量和尺度等统计方法，对英国、法国和荷兰三国的国情国力进行了系统的数量对比分析，从而为统计学的形成和发展奠定了方法论基础。因此，马克思曾说："威廉·配第——政治经济学之父，在某种程度上也是统计学的创始人。"

政治算术学派的另一个代表人物是约翰·格朗特，如图 0.2 所示。

图 0.2　约翰·格朗特

格朗特出生于伦敦，其父母经营一家服装店，他从小在店里帮工，受到了良好的英语教育。小格朗特是一个勤奋的孩子，每天在店铺开门前，他都会自学法文和拉丁文。他以 1604 年伦敦教会每周发表一次的“死亡公报”为研究资料，于 1662 年发表了名为《关于死亡公报的自然和政治观察》的论著。在论著中，他分析了 60 年来伦敦居民死亡的原因及人口变动的关系，首次提出通过大量观察，可以发现新生儿性别比例具有稳定性和不同死因的比例等人口规律，并且第一次编制了“生命表”，对死亡率与人口寿命进行了分析，在当时的学术界获得很高的评价。随后，他被英国皇家学会收为会员。他的研究清楚地表明了统计学作为国家管理工具的重要作用。

政治算术学派主张用大量观察和数量分析等方法对社会经济现象进行研究，为统计学的发展开辟了广阔的前景。

（2）国势学派——“统计学”的命名者。

国势学派诞生于 17 世纪的德国，由于该学派主要以文字记述国家的显著事项，所以又被称为记述学派。戈特弗里德·阿亨瓦尔和赫尔曼·康令是该学派的代表人物。

康令和阿亨瓦尔都在德国大学开设了相关课程来讲授政治活动家应具备的知识。阿亨瓦尔在其主要著作《近代欧洲各国国势学纲要》中讲述了“一国或多数国家的显著事项”，主要用对比分析的方法研究了国家组织、领土、人口、资源财富和国情国力，比较了各国实力的强弱，为德国的君主政体服务。

该学派在进行国势比较分析中，偏重事物性质的解释，而不注重数量对比和数量计算，但却为统计学的发展奠定了经济理论基础。

2．19 世纪——社会统计学派 VS 数理统计学派

从 18 世纪开始，统计学进入了飞速发展阶段。到了 19 世纪，各学派的主要学术观点已成型，这个阶段涌现出来的学派可以说是政治算术派和国势学派的融合与衍生。

（1）数理统计学派——理论在争论中前进。

要说这个派系，不得不提概率论。16 世纪 20 年代，有个酷爱赌博、算命、开方子的意大利数学家卡尔达诺，根据长期的赌博经验，计算了概率；17 世纪，意大利的伽利略通过对赌博问题的研究，创立了早期的概率理论；17 世纪下半叶，瑞士数学家雅克布·伯努利发现大数定律中最早的一个定理——伯努利大数定理；19 世纪初，

法国的拉普拉斯终于集古典概率之大成，初步奠定了数理统计学的基础。古典概率理论的日趋成熟，促使统计科学开始酝酿着嬗变。

19 世纪中叶，比利时人阿道夫·凯特勒把概率论引进统计学，进而形成数理统计学派。在学科性质上，凯特勒认为，**统计学是一门既研究社会现象又研究自然现象的方法论科学**。在当时，这一思想已属突破性的创举，它已经让统计学在准确化道路上跨进了一大步，为数理统计学的形成与发展奠定了基础。19 世纪中叶到 20 世纪中叶，数理统计学得到迅速发展：英国生物学家高尔顿提出并阐述了相关的概念；K·皮尔逊提出了标准差、卡方检验等方法；戈塞特建立了“小样本理论”；费希尔在样本相关系数的分布、方差分析、实验设计等方面的研究中做出了重要贡献。到了 20 世纪中期，数理统计学的基本框架已经形成，统计学也逐渐从记述性统计转变为推断性统计。数理统计学派已然成为英、美等国统计学界的主流。

（2）社会统计学派——重“质”的学派。

社会统计学派诞生于 19 世纪后半叶，创始人是德国的克尼斯，主要代表人物有恩格尔、梅尔等人。他们融合了国势学派与政治算术学派的观点，沿着凯特勒的“基本统计理论”向前发展，但在学科性质上认为统计学是一门社会科学，是**研究社会现象变动原因和规律性的实质性科学**，以此同数理统计学派的通用方法相对立。

社会统计学派在研究对象上认为，统计学研究总体而非个别现象，而且认为由于社会现象的复杂性和整体性，必须对总体进行大量观察和分析，研究其内在联系，才能揭示现象的内在规律。这是社会

统计学派的“实质性科学”的显著特点。

随着社会经济的发展，要求统计学提供更多的统计方法；社会科学本身也不断地向细分化和定量化发展，也要求统计学提供更有效的调查整理、分析资料的方法。因此，社会统计学派日益重视方法论的研究，出现了向实质性方法论转化的趋势。不过，社会统计学派和数理统计学派的对立点建立在对“质”和“量”的争论上。社会统计学派仍然强调在统计研究中必须以事物的“质”为前提和认识事物“质”的重要性，而数理统计学派则侧重计“量”不计“质”的方法论。

在 20 世纪以前，统计学的研究领域主要包括人口统计、生命统计、社会统计和经济统计。随着社会、经济和科学技术等多领域的共同发展，如今统计学的范畴已覆盖了我们社会生活的一切领域，成为通用的方法论科学。特别是第二次世界大战以来，由于经济、社会、军事等方面的客观需要，统计预测和统计决策科学有了很大发展，使统计走出了传统领域而被赋予新的意义和使命。在近阶段的统计学发展史上，贝叶斯派系的统计学获得越来越多学者的关注，也是推动近代统计学发展的新动力。

贝叶斯统计学派的主导思想来源于贝叶斯的后验概率，它和之前所说的各大派系（一般统称为经典统计学派）的区别在于是否利用先验信息。贝叶斯统计学派认为，利用这些先验信息不仅可以减少样本容量，而且在很多情况下可以提高统计精度；而经典统计学派则忽略了这些信息。

诚然，贝叶斯统计学派与经典统计学派有着较大区别，但是它们各有优缺点，各有其适用范围。经典统计学派历经了时间的冲刷，理

论体系已然相当成熟；而贝叶斯统计学派带来的新理念，势必会激起新的一轮方法论研究。两种方法相辅相成，在很多情况下，二者得出的结论在形式上是相同的，在结果上也具有同质性。

我们来梳理一下 20 世纪统计主要理论的发展，如图 0.3 所示。

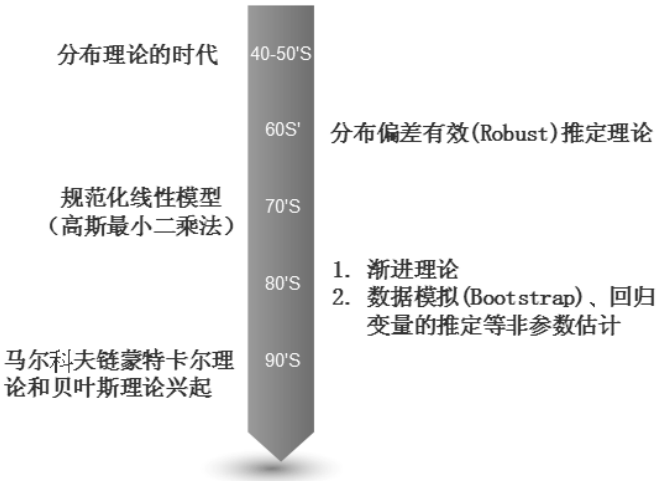


图 0.3　20 世纪统计主要理论时间轴

随着计算机等信息化工具的普及，统计学也具备了普及的条件，统计思维必将成为现代人的必备思维之一。

第 1 章

你的数据从何而来

前面我们对“统计”一词有了一个粗略的定义：对数据进行收集和整理，并在此基础上加以分析和做出科学决策。既然统计的主体是数据，那么问题来了：数据从哪里来？得来的数据可信吗？我们怎样才能获得高可靠度的数据呢？本章要回答的就是这三个问题。首先让我们来看看数据可以从哪些方面获得。

如果对进行统计分析所使用的数据做大致分类，可以将其分为两类：一手数据和二手数据。

什么是一手数据？打个比方，就好比新房一样，一手数据是刚刚“建造出来”的，也被称为原始数据。一手数据可以分为调查、观察所得数据和实验所得数据。比如调查取样时获得的数据；又比如通过化学、物理实验得出的各种数据。

那么什么是二手数据呢？同样作个类比，二手数据就像二手房一样，是经过“转手”的，常见的二手数据是利用文献、统计年报、行

业协会信息及数据库等统计好的数据资料。

一手数据和二手数据各有优缺点：

- 一手数据能够提供量身定制的信息。比如你需要做哪项研究，即可专门为此设立调查问卷，获得最为直接、最为相关的数据，便于进行有针对性的分析研究。不过搜集数据需要较长的时间，而且花费的成本也更多，最重要的是在搜集数据的过程中，采用何种调查方法对结果具有重要影响。

- 二手数据通常能够廉价，甚至免费获得，而且可以在更短的时间内进行分析。不过在采用二手数据之前，我们必须考虑这些数据的含义是什么、它的获取方法和计算口径是什么、数据的可靠度如何、数据是否具有可比性等问题。如果数据获取的初始目的与研究目的不相关，那么还需要进一步梳理信息来提取内容。

相较于二手数据，我们往往会将数据来源的质量焦点更多地集中在原始数据上。这就涉及怎样才能获得高可靠度的原始数据问题。

通常而言，采集原始数据的方法主要有**普查**、**抽样调查**和**实验观察**，接下来我们主要探讨一下普查和抽样调查的基础知识。

1.1 “不可能完成的任务”——普查

普查对于老百姓而言并不陌生，最为熟知的就是我国每十年都会进行一次的全国人口普查。对于普查，其种类不仅限于人口，还有每逢“3”的年份进行第三产业普查，每逢“5”的年份进行工业普查，每逢“7”的年份进行农业普查，每逢“1”或“6”的年份进行统计

基本单位普查。看到这里，你可能会产生疑问：为什么普查的时间跨度那么大，而且往往由国家、政府牵头？这就引出了普查的概念和特点。

简单说来，普查类似于企业定期的盘点工作：在某个时点、在某个范围内对账款货物进行清点。专业描述为：普查是为了某种特定的目的而专门组织的一次性的全面调查。在通常情况下，它调查的是在一定时点上的社会经济现象的总量，但也可以用来调查某些非总量的指标。

普查具有以下几个特点：

（1）一次性或周期性。由于普查耗费的人力、物力、财力都是巨大的，因而不可能经常开展。一般进行常规性的普查都会有一定的周期性，比较常见的周期为 5 年或 10 年。设定周期还有一个好处，就是便于数据的利用，有了规律的周期，可以更方便地进行数据比较。

（2）统一的标准时点。所谓标准时点，就是规定一个时间点，无论普查员登记在哪一天进行，登记的指标都是反映那个时间点上的情况。

为什么要规定一个时间点？因为普查的开展往往有一个期限，在这段时间内，万事万物都可能发生改变，为了避免调查时因情况变动而产生重复登记或遗漏现象，所以必须规定一个时间点。

（3）统一的普查期限。虽然普查工作繁复，但总不可能无限期地进行。开展普查时，在普查范围内，各调查点应该尽可能地同时登记，力求在最短的期限内完成，以便在方法和步调上保持一致，保证资料的准确性和时效性。

（4）规定的普查项目和指标。普查时必须按照统一规定的项目和指标进行登记。对项目和指标进行规定是为了避免影响汇总和综合，降低资料质量。需要强调的是，在指标的计算和解释上也要保持一致，以便进行历次调查资料的对比分析和观察社会经济现象的发展变化情况。

（5）基础性和有限性。相对而言，普查获得的数据是比较准确和规范的，所以可以利用其为抽样调查或其他调查提供基本依据；不过，以客观来看，因为普查的适用范围较窄，且时间跨度较长，所以只能调查一些最基本及特定的现象，这也是普查的局限性。

1.2 “四两拨千斤”——事半功倍的抽样调查

相较于普查的声势浩大，抽样调查就要低调多了。因为抽样调查涉及的调查对象相对规模小，且方法灵活，所以在日常生活中，这种数据取样方法的使用率最高。

说得严谨一些，抽样调查可以这样理解：它是从研究对象的全部单位中抽取一部分单位进行考察和分析，并用这部分单位的数量特征去推断总体的数量特征的一种调查方法。其中，被研究对象的全部单位称为“总体”；从总体中抽取出来，实际进行调查研究的那部分对象所构成的群体称为“样本”。

抽样调查有很多优点，比如节省费用的同时可以提高效率（费用和调查精度也密切相关，如何做到二者的均衡是一门艺术）；又比如可以快速、准确地得到信息（利用概率论和统计学的相关知识，可以方便、准确地从样本出发推算总体的参数情况）。

但是抽样调查是一件有概率参与其中的调查工作，所以伴随而来

的就是抽样的误差。误差其实不可避免，控制误差也是整个抽样方案设计时需要着重思考的地方。不过对于误差，如果采取科学的抽样方法，是可以做到有效控制和规避的。

要说抽样界里哪个抽样案例最著名，那就不得不提盖洛普在1936年的那次总统竞选上成功预测罗斯福获胜之例了。

1936年，美国的总统选举进入白热化阶段。在选举进行的同时，《文学摘要》杂志和盖洛普舆论研究所等三家民意调查机构就对谁会成为本次选举的最后赢家分别做了预测。当时盖洛普使用了定额抽样法，根据调查对象的年龄、性别、受教育程度等在全国按比例选择调查对象，抽取了大约5万名民众就得出了罗斯福会取胜。而《文学摘要》杂志花费了大量的人力、财力、物力，采用了大规模的模拟选举。他们以电话簿上的地址和俱乐部成员名单上的地址发出 1000 万封信，收到回信200万封。这种大规模的调查在调查史上都是少见的，因而杂志社坚信自己的调查统计结果——兰登会以57%的比例获胜，并为此进行了大力宣传。现实是残酷的，再多的资源耗费都不及科学的调查方法，最后罗斯福以62%的巨大优势获胜，连任总统。

在生活中，我们经常看到一些调查结果和我们自身感觉差异较大。那么，怎样做才能使通过抽样调查得到的数据，经过分析之后可以得出令人信服的结果呢？我们不妨来总结一下1936年那次总统选举预测的经验教训：

盖洛普舆论研究所在选择样本的时候对调查对象做了定额抽样，其实质并非是随机的，在具体操作上可能存在一定的主观选择性，在选择对象的时候考虑到了性别、年龄和政治观点等因素，这在某种程度上弥补了非随机抽样的不足，使得所获样本更具有社会代表性。相

比之下，《文学摘要》杂志的样本不是从总体（全体美国公民）中随机地抽取。1936 年，美国有私人电话和参加俱乐部的家庭都是比较富裕的家庭。1929—1933 年间发生了世界经济危机，这使美国经济遭到沉重打击。“罗斯福新政”动用行政手段干预市场经济，损害了部分富人的利益，但广大的美国公民却从中得到了好处。所以，从这部分富人中抽取的样本严重偏离了总体，导致样本不具有代表性。

这个故事揭露了一个很实用的规律：**当样本的选取方法发生偏差时，你有再多的样本都是徒劳的，这只会让你在错误的道路上走得更远点罢了。**

那么，怎样设计抽样方案才能既省时又省力，还能得到接近总体的调查样本呢？下面来看看几种常用的抽样方法。

1．简单随机抽样——随机≠随意

简单随机抽样法是所有抽样方法中最简单也最为基础的一种方法，它是等概率抽样方法。它的抽样理念是从总体中选出抽样单位，从总体中抽取的每个可能样本均**有同等被抽中的概率**。

简单随机抽样是日常生活中常用的，比如公司年会上的抽奖活动，就会采用随机抽签的方式来选出那个幸运儿。不过，有时候，我们也会无意间让随机抽样沦为随意抽样。仍以年会抽奖为例，如果所用的摇奖箱没有将奖券混匀，当参会者依此投入自己的奖券时，到得越早的被抽中的概率越小，因为在抽签时人们更倾向于抽取中段部分的奖券。

另一种将随机抽样沦为随意抽样的情况就是，在抽样时完全按照抽样者的主观意愿，随意进行抽样对象的选择，这在街头的抽样调查

中经常见到。比如，一家化妆品公司派出几名调查员去了解市场对化妆品的认知度情况，有的调查员嫌麻烦，就在工作日对自家小区进行了采样，这样的采样效果就会大打折扣。因为工作日在住宅小区里闲逛的往往是退休人员，年轻的职业女性都在职场上拼搏，而化妆品的消费人群又以职业女性为主，这就是典型的随意抽样。

为了让获得的样本更符合“随机”而非“随便”，使它更能代表总体，可以通过科学的方法来实施抽样。具体操作方法为：在抽样时，将抽样总体中的抽样单位用 1～N 编码，然后利用随机数码表、抽签法或专用的计算机程序确定处于 1～N 的随机数码，那些在总体中与随机数吻合的单位便成为随机抽样的样本。举个简单的例子，如果要对某工厂车间生产的产品进行质量抽查，采用简单随机抽样法，首先将产品依次编号；然后根据要抽取的样本量来选择是用抽签法还是用随机数发生器来决定抽取的编号；接下来，只需取出与编号相对应的产品，对其进行质量检查即可。

简单随机抽样方法虽然简单，进行误差分析也比较容易，但当总体的容量非常大时，该方法既费时又费力，因此它适合总体容量较小、个体之间差异较小，并且可以让样本等概率入选的抽样情况。

2．系统抽样——机械→效率

所谓系统抽样，其实并没有什么复杂的系统，换个名字就能很好地理解，如等距抽样。它和随机抽样是近亲，近到什么程度？先来看看什么是系统抽样。

系统抽样是将总体中的各单位按一定顺序排列编号，根据样本容量要求确定抽选间隔，然后随机确定起点，每隔一定的间隔抽取一个单位的一种抽样方式。

具体做法如下：首先将总体从 1～N 相继编号，并计算抽样距离 $K=N/n$（式中，N 为总体单位总数，n 为样本容量）；然后在 1～K 中抽取一个随机数 k_1，作为样本的第一个单位，接着取 k_1+K、k_1+2K……（这个过程其实就构成了一个系统），直至抽完 n 个单位为止。它和随机抽样的区别在于，后者通过抽签或随机数来获得编号，前者通过固定起点和距离来获得编号。

系统抽样法简单方便、经济有效，在很多时候它是随机抽样的优秀替代品，得到的样本与简单随机抽样得到的样本几乎相同。不过系统抽样也有随机抽样所不具备的缺点，那就是它抽取出的对象在总体中是均匀分布的，比如都间隔 10 个单位，这就需要我们对总体结构有一定的了解。如果能充分利用已有信息对总体单位进行排序后再抽样，则可提高抽样效率。

3．分层抽样——偏心 or 公平

针对简单随机抽样适合容量较小、个体差异小的总体，分层抽样法是很好的一种补充抽样方法。分层抽样法是根据某些特定的特征，将总体分为同质、不相互重叠的若干层，再从各层中独立抽取样本。与简单随机抽样不同的是，它是一种不等概率抽样。

不等概率抽样其实就是在抽样时有目的地设置不同的权重。以全国 1%的人口抽样为例，如果使用随机抽样或者系统抽样法，对于有十多亿人口的中国来说，要抽取 1000 多万人口作为样本，那么仅编号就是一项庞大的工程，更别说因此抽出的样本中会出现遗漏问题了。但如果根据地区或民族人口分布比例来制定抽样比，那么抽样调查工作不仅能有条理地开展，调查工作量也会分散开，获得的样本也会更具有代表性，这样的抽样可操作性更高、效果更好。

分层抽样法的特点是：**利用辅助信息分层，各层内差异小且具有同质性，但各层间差异尽可能大。这样的分层抽样能够提高样本的代表性、总体估计值的精度和抽样方案的效率。**

但分层抽样与简单随机抽样也有区别。如果从相同的总体中抽取两个样本，一个是分层样本，另一个是简单随机抽样样本，那么相对来说，分层样本的误差更小一些。反过来，当我们确定了抽样误差水平后，那么更小的分层样本将达到这一目标。

不过分层抽样的抽样框比较复杂，所需的费用较高，在计算和分析误差时也会较为复杂。通常在遇到总体情况复杂、个体之间差异较大、总体数量较多的情况时会选择这种抽样方法。

4．群抽样——普查迷你版

之前说过，普查是一项耗费巨大的“工程项目”，虽然能够了解总体的信息，但是代价也很大。其实，在抽样调查里也有一种被称为“小普查”的方法——整群抽样。

如果总体可以分为 N 个初级单位（它们就是群的概念），每个群包含若干个体，通过某种方式（常用的有随机抽样方式）从总体中抽取 n 个群，然后对这些群中的所有个体进行普查（这就是小普查的由来），则称为整群抽样。

用一句话来概括，**整群抽样其实不是直接抽个体样本，而是抽群。**还是通过例子来理解这种方法。

比如，想要了解某市中学生近视发病率，如果采用整群抽样法，则可以这样操作：该市共有 48 所中学，这 48 所中学构成了调查的总体。对于调查者来说，只需抽取 48 所中学中的几所，然后对抽中的

学校进行学生的普查即可。为什么抽取学校而不直接抽取学生呢？因为从抽样工作开展上来说，直接抽取学校更为方便，而学校与学校之间对于学生近视发病率而言并没有太多特征上的差异；相对的，各学校中因为学生的年级、性别不同，学生个体是存在差异性的，对学校中的学生进行普查则可以了解学生发生近视的特征性。

整群抽样和分层抽样之间的区别还是很明显的：分层抽样要求层与层之间差异要大，同一层内的个体差异要尽量小（因为分层是对抽个体的辅助行为，**最终是抽个体**）；而整群抽样则要求群与群之间差异要小，而群内个体差异越大越好。

整群抽样的优点在于样本比较集中，可以降低调查费用，便于组织，但得出的结果误差较大。比较上述 4 种抽样方法的抽样误差，通常情况下：整群抽样 ≥简单随机抽样 ≥系统抽样 ≥分层抽样。

5．多阶段抽样——终极大成者

如果能掌握上述 4 种抽样方法，则能完成大多数的抽样调查。不过有些调查的复杂程度和工作量是超乎想象的，比如，若想单一地使用上述任何一种方法来完成全国 1%人口抽样这种规模的调查，则收效甚微，这时就需要进行抽样方法的综合，进行二阶段甚至多阶段的抽样。

多阶段抽样法其实不是一种具体的抽样方法，而是一种抽样组合法，比如分层抽样和整群抽样的结合。以二阶段抽样为例，从总体上所有一阶单元中抽取一部分单元，相当于从总体所有群中抽取部分群的整群抽样；而在每个抽中的一阶单元中分别抽取部分二阶单元，就相当于分层抽样。即先整群，后分层。因此，二阶抽样从技术上看是整群抽样与分层抽样的综合。抽样形式对比如表 1.1 所示。

表 1.1　抽样形式对比

抽样形式	第一阶段	第二阶段
分层抽样	抽全部	抽部分
整群抽样	抽部分	抽全部
二阶抽样	抽部分	抽部分

仍以 1%人口抽样为例，对于落实到区级的人口抽样方案，就可以这样考虑：先在某行政区每个街道抽取 n 个居民小区，再对 n 个小区根据门牌号进行系统抽样或者随机抽样。这就是一个多阶段抽样的应用。

研究数据的来源其实是一个很有趣的话题，不同的数据获取方法会获得不同的调查结果，是花力气普查得到最原始、最全面的数据，还是用点“小伎俩”抽取一些样本数据来推算总体数据，这不仅需要专业知识，还需要丰富的“实战”经验。

不管你掌握了多少种抽样方法，笔者并没有在书中探讨各类抽样所需的样本量及抽样会产生的误差计算这类专业话题，只希望大家能发现抽样的美，在可能的情况下多尝试几种不同的抽样方法，然后择优选择。

☆本章重点归纳

- 数据来源分类
  - 一手数据：也称原始数据，指通过人员访谈、询问、问卷、测定等方式直接获得的数据 ⇨ 优点：时效性和相关性强
  - 二手数据：利用文献、统计年报及数据库等前人统计好的数据资料 ⇨ 优点：获取成本低，且现成可用。一般可以长时间保存，方便生成数据趋势图

- 数据来源方法优缺点对比：

	普查	抽样调查			
		简单随机抽样	系统抽样	分层抽样	整群抽样
优点	1．获得全面资料 2．准确性高	1．操作简单 2.均数及相应的标准误差计算简单	1．易于理解 2．简便易行	1．样本代表性好 2．抽样误差减少	1．便于组织 2．节省经费
缺点	1．工作量大 2．花费大 3．组织工作复杂 4．调查内容有限 5．易产生重复和遗漏	总体较大时难以编号	总体有周期或增减趋势时，易产生偏差	抽样过程繁杂	抽样误差大于单纯随机抽样

第 2 章

掌握指标学会数据分析

如果你掌握了均值、方差（标准差）、峰度、偏度这几个指标，就能对数据进行分析。不信？且看下文分解。

2.1 被误解还是“被平均”

在正式开始本章的知识点介绍之前，我们先来看一篇新闻报道。注意，我们所关注的并不是新闻报道的时间，而是报道中用到的几个词。

“2014 年，发改委官员曾表示，我国**人均 GDP** 已达到 6700 多美元，属于中高收入国家的行列。目标是希望通过“十三五”的努力，用世界银行的标准接近高收入国家的行列。”

这则新闻报道其实说的并无不妥之处，按照理论来说，如果中国能保持目前的发展速度，那么 10 年左右进入高收入国家行列是顺理成章的事。到 2020 年，中国人均 GDP 达到 1 万美元也不是梦想。但

很多人还是质疑自己可能“被高收入”了。

其实，我国民众对统计数据的“不适”已经不是第一次了，近年来，网络吐槽“被平均”、“被幸福”等情况屡屡出现。比如 2012 年，某大学发布的《中国民生发展报告 2012》中提及，全国家庭的平均住房面积为 116.4 平方米。这个结论是不是让你很诧异？那么，到底是什么原因导致统计结论让人感觉与自身情况不符呢？

抛开理性，你会发现这种感觉其实很好理解。对于广大人民群众而言，要判断统计数据是否真实，最好的印证和参照物就是自身和周围的生活状况。如果你发现自己及周围人的情况和统计结论有不小的出入，那么感觉“被平均”就再自然不过了。但是如果仅凭统计数据和自身感受不一致就认为数据不正确，那就比较片面了。

引起误解的还有一个很巧妙的用词——“人均”。这一平均，很多数据就被“削峰填谷”、加权计算了，呈现在你眼前的是一个总体性指标，作为个体的你只能略作参考，它和个体数据还是有很大差异的。

下面讲解本章的第一个重要知识点——平均数。先来看一道题。

假设有 100 人，他们的平均身高为 163.5cm，请判断以下三句话的对错：

（1）身高低于和高于 163.5cm 的约各有 50 人。

（2）全部人员的身高加起来共 16350cm。

（3）每 10cm 分成一组，160～170cm 的人数是最多的。

在公布答案之前，先来看看这三句话分别涉及哪些概念。

- “他们的平均身高为 163.5cm”——平均数（算术平均数）。
- “身高低于和高于 163.5cm 的约各有 50 人”——中位数。
- “每 10cm 分成一组，160～170cm 的人数是最多的”——众数。

在统计学上把平均数分为两大类：数值平均数和位置平均数。前者包括算术平均数、加权平均数和几何平均数，后者包括中位数和众数。这几个指标通常用来描述总体均值情况，但它们是不是真的那么平均？要正确理解它们，还得回到指标的本质含义来探讨。

2.1.1 数值平均数——最熟悉的陌生人

数值平均数可以说是最为熟悉、最为常用的表示平均的指标。数值平均数可以分为好几类，这里仅对算术平均数、几何平均数和调和平均数进行简单介绍。

1．算术平均数

算术平均数通常也称为均值，可分为简单算术平均数和加权算术平均数两类。在实际生活中，并不是每次计算均值时，各项都拥有相同的权重（相同权重时，称之为简单算术平均数），当各项权重不相等时，计算平均数时就要采用加权算术平均数。

一般简单算术平均数可以通过如下公式得到：

$$\frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

而加权算术平均数则是把原始数据按照合理的比例来计算。若在 n 个数中，x_1 出现 f_1 次，x_2 出现 f_2 次，…，x_n 出现 f_n 次，那么加权平均数的公式可以如此推导：

$$\frac{x_1f_1+x_2f_2+x_3f_3+\cdots+x_nf_n}{f_1+f_2+f_3+\cdots+f_n}$$

式中，$f_1, f_2, \cdots, f_n$ 是 $x_1, x_2, \cdots, x_n$ 的权。

为了更好地理解，我们来看一个简单的例子。某人特别爱吃青菜，于是某个周日决定去买点青菜亲自下厨。当他兴冲冲地来到菜市场时，发现在甲摊位青菜卖 2 元/斤，而在乙摊位青菜卖 3 元/斤。由于不知道到底哪个摊位的菜更好，他决定从甲、乙两个摊位各购买 1 斤，求平均价格。

这种情况很简单，可直接用简单算术平均数的公式求得平均价格为：(2+3)/(1+1)=2.5（元/斤）。

现在假定其他条件不变，若从甲摊位购买 2 斤，从乙摊位购买 1 斤，再来求平均价格。

加权算术平均数=(2×2+1×3)/(2+1)=2.3（元/斤）。

在这个例子中，我们所选用的是同一种蔬菜，具有同质性。但在运用算术平均数的时候往往忽略了这个内涵要求，从而导致结果有失偏颇。比如，在电梯里，你的体重是 120 斤，有个小孩的体重是 80 斤，还有一个箱子重 400 斤，平均重量是(120+80+400)/3=200（斤）。这时，能说三者的平均重量是 200 斤吗？这个均值只能说明电梯负重了多少，此时的平均重量并没有什么参考意义。

算术平均数虽然计算简单、理解方便，但它有一个致命的缺点——容易受到异常值的影响。

请看下列数字：5、7、5、4、6、7、8、5、4、7、8、6、20，其

平均值为 7.1，实际上大部分数据（有 10 个）不超过 7，如果去掉 20，则剩下的 12 个数的平均数为 6。之所以算术平均数容易受到异常值的影响，是因为它反应灵敏，每个数据或大或小的变化都会影响最终结果。

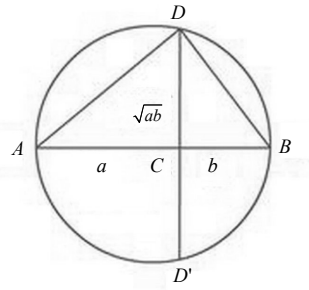
2．几何平均数

比起众所周知的算术平均数，几何平均数就显得有点小众，但是几何平均数有着无可替代的地位。既然取名为几何平均数，那么它自然是具有几何意义的。可是，一个平均数怎么会和几何有关？其实在中国古代数学书中提到矩形面积时，往往就是用长、宽的几何平均数来表示的。我们来看看到底什么是几何平均数。

几何平均数是指 n 个观察值连乘积的 n 次方根，公式如下：

$$\sqrt[n]{x_1 \bullet x_2 \bullet x_3 \bullet \cdots \bullet x_n}$$

仅有公式，还是没有看到它的“几何”在哪。别急，先来看下面这张图：



所谓几何关系，可以这样理解：过一个圆的直径上任意一点作垂线，直径被分开的两部分为 a,b，那么这条垂线在圆内的一半长度就是 $\sqrt{ab}$，并且$(a+b)/2 \geqslant \sqrt{ab}$。这就是它的几何意义。一般来说，几何

平均数主要用于以下几个方面：

- 用来对比率、指数等进行平均。
- 用来计算平均发展速度。
- 用来计算复利下的平均年利率。

下面来看一个小案例（引自网络博客）：现在有两只基金投资组合，投资了 4 只股票，盈亏率情况如下：

组合方案 A：+10%，−10%，10%，−10%

组合方案 B：+30%，−30%，30%，−30%

如果让你选择一只基金投资组合，你认为哪只基金盈亏比较平衡呢？先用简单的算术平均数来比较一下：方案 A 和方案 B 的盈亏都是 0，甚至你会认为方案 B 更好些，符合“挣得多，赔得多”的风险投资理念。

但如果采用几何平均数再进行计算：

组合方案 A：$(\sqrt[4]{1.10\times0.90\times1.10\times0.90}-1)\times100\%$，得出平均约有 0.5%的亏损。

组合方案 B：$(\sqrt[4]{1.30\times0.70\times1.30\times0.70}-1)\times100\%$，得出平均约有 4.6%的亏损。

可以看出，两只基金投资组合都是亏损的，但如果必须选择一只基金投资组合，则方案 A 比较稳妥。这个案例是不是让你对几何平均数的优势有了深刻印象？

不过几何平均数也有自己的不足，在变量值可能出现负数的情况下，不能用样本的连乘积或者几何平均值，因为变量的负值会带来连乘积的值时正时负。所以对于变量可能存在负值的样本（如摄氏气温），不能统计其几何平均值。类似地，变量可能为 0 的样本，会使连乘积等于 0，所以这类变量也不能统计几何平均值。

3．调和平均数

调和平均数也叫倒数平均数，是总体各统计变量倒数的算术平均数的倒数。在数学中，调和平均数与算术平均数都是独立的、自成体系的，计算结果前者恒小于等于后者。但统计加权调和平均数则与之不同，它是加权算术平均数的变形，附属于算术平均数，不能自成体系，且计算结果与加权算术平均数完全相等。具体公式如下：

$$\frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}+\cdots+\frac{1}{x_n}}$$

调和平均数主要用来解决在无法掌握总体单位数（频数）的情况下，只有每组的变量值和相应的标志总量，而需要求得平均数的问题。

那么，什么时候可以用调和平均数进行计算呢？

其实，调和平均数不被熟知的一个重要原因是其应用的范围较小。在实际中，往往由于缺乏总体单位数的资料而不能直接计算算术平均数，这时就需要用调和平均法来求得平均数。

通常在遇到需要计算平均速度（一般指物理中速度、位移的解题）、平均利润率、平均成本等指标时可以使用调和平均数。不过，即便它和算术平均数关系紧密，二者也不能混用。调和平均数具有以下特征：

- 调和平均数易受极端值的影响，且受极小值的影响比受极大值的影响更大：上端值越大，平均数向上偏离集中趋势就越大；反之，下端值越大，平均数向下偏离集中趋势越大。
- 只要有一个标志值为 0，就不能计算调和平均数（分母不能为 0）。
- 当组距数列有开口组时，其组中值即使按相邻组距计算，假定性也很大，这时的调和平均数的代表性很不可靠。

综上，不同的数值平均数有着不同的适用范围：算术平均数适用于简单且较直观地表现中心位置；当数据呈倍数关系或不对称分布时（增长率或生长率、动态发展速度），适合使用几何平均数；调和平均数适用于观测值是阶段性变异的资料。其数值大小排序为：调和平均数≤几何平均数≤算术平均数。

2.1.2 位置平均数——关键的排序

如果非要用一个词来区分位置平均数和数值平均数的区别，则可以用“次序”一词。在计算数值平均数的时候，一般不会刻意地对数据进行从小到大的排序，而是直接将数值和权数一并放入算式中，计算得出一个平均数。但是位置平均数则完全不同，不同到有时只需从小到大排序，或者把每个数值出现的次数从少到多排序，无须计算就可以得到一个均值。下面来看看两个位置平均数的代表：中位数和众数。

1．中位数

中位数是中间位置的数字。中位数将所有的观察值一分为二，一半的数字比它大，另一半的数字比它小。那么，现实问题中如何求得

中位数？在需要求得中位数时，首先需要把所有的观察值从小到大进行排序。

举个例子：小时候，老师最喜欢在考完试后进行排名，这就是一个排序过程。如果该班级共有 51 名学生，那么考试成绩从最低分（或最高分）开始依次排序，直至最高分（或最低分），这样就会得到一组递增（或递减）的数据。51 名学生正好第 26 名是中间者，我们就选他的考分作为考试成绩的中位数。但如果有 52 名学生呢？如果把人数一分为二，排名第 26、27 位的两名学生均处在中位，该怎么选？可见，求中位数有一个注意点，那就是观察值的个数是奇还是偶。如果观察值的个数是奇数，那么求适中的数值即可；如果观察值的个数是偶数，那么通常取最中间的两个数值的算术平均数作为中位数。用公式描述如下。

若有观察值 $x_1,x_2,\cdots,x_n$，若 n 为奇数，则中位数为

$$m_{0.5}=x_{(\frac{n+1}{2})}$$

若 n 为偶数，则中位数为

$$m_{0.5}=\frac{x_{\frac{n}{2}}+x_{\frac{n}{2}+1}}{2}$$

公式很简单，理解也不难，但什么时候适合用中位数呢？若要回答这个问题，就要回到本章开头所提的那个问题：你是不是总觉得自己在各种数据面前有“被平均”之感？

举个例子：2014 年全国平均工资为 4.99 万元，月平均工资为 4000 多元，这也就罢了；北京的平均工资达 77 560 元，月平均工资为 6463

元，这就很让人艳羡了；尤其是，全市城镇非私营单位就业人员年平均工资为 102 268 元，月均达 8522 元——对于大部分人而言，岂不是“拖后腿”、“被平均”？了解了算术平均数你会发现，公布的数据可能并没有问题，问题在于工资收入的分布是否适合使用算术平均数来表示均值？我们来看看工资的大致分布图，如图 2.1 所示。

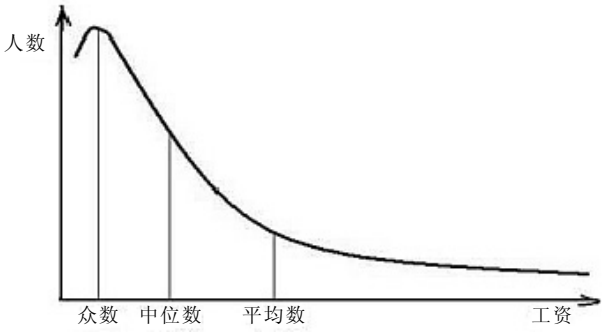


图 2.1 工资收入均值分布图

从图 2.1 中可以看出，一般来说，一个人群中的工资收入分布，众数往往偏左，而平均数往往偏右。这说明低收入人群占多数，高收入人群占少数，工资收入呈偏态分布。而从报道中的数据可以推测，导致平均工资如此高的原因是那部分人数少但收入高的人群拉高了均值。

所以，一般情况下，对于收入、房价等数据，在公布算术平均数的同时需要公布中位数作为参考，这样的数据会更有实际意义。那么，既然知道中位数具有很强的参考辅助功能，收入为何不能公布一个中位数呢？这又回到了排序问题。如果我们所拥有的观察值是有限个数且容易操作点数排序的，那么一切都可以顺利进行。但是如果要得到一个收入的中位数，那就意味着需要对该城市（该国）所有工作者都进行调查（而事实是我们的工资收入都是以企业为单位进行统计的），

工作量实在太大，排序就显得尤为困难。

而中国香港在收入统计方面不仅公布了平均数(一般指算术平均数)，还公布了百分位数和中位数，如图 2.2 所示。

	2013 年 5 月至 6 月 May – Jun 2013	2012 年 5 月至 6 月 May – Jun 2012	增减百分率 Percentage change
每月工資分布（港元）：所有僱員 **Monthly wage distribution (HK$) :** **All employees**			
第十個百分位數 10th percentile	7,700	7,300	+6.0
第二十五個百分位數 25th percentile	10,000	9,500	+5.8
第五十個百分位數 50th percentile	14,100	13,400	+5.2
第七十五個百分位數 75th percentile	22,000	20,900	+5.3
第九十個百分位數 90th percentile	36,200	35,800	+1.2
按性別劃分的每月工資中位數（港元）： **所有僱員** **Median monthly wage by sex (HK$) :** **All employees**			
男 Male	15,800	15,000	+5.4
女 Female	12,200	11,700	+4.3

图 2.2　中国香港工资分位数分布图

由图 2.2 可以引入一个新的知识点——四分位数和五数概括法。

先来说说四分位数：四分位数中有一个分位数我们已经认识了，它就是中位数，在四分位数中排行第二，代表数值由小到大排列后第 50%的数字；其余分别为第一四分位数（Q1），又称“较小四分位数”，等于观察值中所有数值由小到大排列后第 25%的数字；第三四分位数（Q3），又称“较大四分位数”，等于观察值中所有数值由小到大排列

后第 75%的数字。所谓的四分位数，其实就是通过三个位置数将数据等量分割成四部分，其中，Q3 到 Q1 之间的距离差又称为四分位距。四分位距越小，说明中间部分的数据越集中；四分位距越大，则意味着中间部分的数据越分散。

五数概括法与四分位数有着紧密关系，五数概括法所用的 5 个数分别为：最小值；第一四分位数（Q1）；中位数（Q2）；第三四分位数（Q3）；最大值。具体的做法也很简单，与求中位数一样，先将数据从小到大排序，然后根据四等分原理获得四分位数，并得到最大值和最小值。

举个例子：有一个观测值样本，内容是 12 个月的月薪数据，按照递增顺序排列如下：

4210 4255 4350 | 4380 4380 4390 | 4420 4440 4450 | 4550 4630 4825

Q1＝4365　　Q2＝4405　　Q3＝4500

（中位数）

根据中位数的计算方法，观察值数据量为偶数，所以在计算四分位数时，需要将 3、4 位，6、7 位和 9、10 位的数据相加除以 2。另外，其中的最小值为 4210，最大值为 4825。因此，上述月薪数据以五数概括为：4210，4365，4405，4500，4825。为什么需要选择这 5 个数？因为通过这样的数据选取，可以对观察值的分布情况有个大致的了解。若这 5 个数之间的间隔比较均匀，那么这 5 个数据具有较好的总体归纳性；若这 5 个数之间的间隔不均匀，那么此时的数据往往不呈正态分布，无论是选择中位数还是算术平均数来描述数据，都有可能出现偏差。

这也就引出了以中位数为代表的位置平均数的一些特点：

- 中位数是以它在所有观察值中所处的位置确定的全体单位的代表值，不受分布数列的极大值或极小值影响，从而在一定程度上提高了中位数对分布数列的代表性。
- 有些离散型变量的单项式数列，当数据分布偏态时，中位数的代表性会受到影响。

2．众数

众数是位置平均数中的另一个重要代表，它将各观察值出现的次数记录下来，选择出现次数最高的观察值作为均值。但是，如果遇到不同的观察值出现同样的次数（且都是最高的）时，怎么办？解决方法是全部命名为众数。所以众数是三大平均数代表中仅有的不唯一取值代表。

比如，对某幅图进行评价，5 位观众分别给出 9 分、7 分、9 分、8 分、6 分。如果用简单算术平均数来计算，则平均分为 7.8 分；若对其进行排序，则为 6,7,8,9,9，中位数为 8；如果通过观察值的出现次数来排序，则 6、7、8 分都出现 1 次，9 分出现 2 次，则众数为 9。可以看出，不同的方法计算出来的均值都不相同。

前面已经大致描述了各类数值平均数的适用范围，也提出了中位数的适用对象，众数也比较挑剔，它往往更适合一些对数值本身不敏感，但对该数值占比有要求的样本。比如，想要了解中国男士的脚码的平均值。如果选用简单算术平均数或者中位数，得出的往往会是 41.25、41.18 这类数据；或者在获得的样本中，排在正中的那位男士的脚码正好是 41 码。这样计算得出的算术平均数不具有实际意义。

这时就要用到众数。比如，我们从某个商场的男士皮鞋销售柜台获得了一份销售明细，其中 38 码的卖了 10 双，39 码的卖了 80 双，40 码的卖了 120 双，41 码的卖了 500 双，42 码的卖了 300 双，43 码的卖了 100 双，44 码的卖了 50 双，共卖出 1160 双。从这份销售明细里可以看出，近半数的男士选择了 41 码的鞋子，说明此码占据市场份额最大，能够代表最多男士的脚码数。从这个例子中可以看到，众数主要用于定类（品质标志）数据的集中趋势的度量。

另外，通过这个案例，看到了不同的平均数之间所具有的区别和联系：只有在所使用的观察值分布呈现偏态（不对称）的情况下，才会出现平均数、中位数和众数的区别。所以，如果观察值呈正态分布，任何统计量都不会出现太大偏差；如果偏态的情况很严重，则可以考虑算术平均数，并辅助参考中位数。

不过，仅仅刻画观察值平均水平的统计量是不够的，统计学中还有刻画数据波动情况和分布形态的统计量。比如，平均数同样是 5，它所代表的数据可能是 1、3、5、7、9，也可能是 4、4.5、5、5.5、6。也就是说，5 所代表的不同组数据的波动情况是不一样的。怎样刻画数据的波动情况呢？统计学中还有方差（标准差）等用来刻画数据离散特征的统计量和偏度、峰度等用来刻画数据分布形态的统计量。结合这些统计量，对观察值的了解才会更深入。

2.2 均值的好朋友——方差（标准差）

在此之前，我们先要对方差（标准差）做个简单的介绍：

方差是在概率论和统计学中用来度量观察值和均值之间偏离程度的指标。样本中各观察值与样本平均数的差的平方和的平均数叫作

样本方差；样本方差的算术平方根叫作样本标准差。当数据分布比较分散（数据在平均数附近波动较大）时，各个观察值与平均数的差的平方和较大，方差就较大；当数据分布比较集中时，各个观察值与平均数的差的平方和较小。因此，方差越大，数据的波动越大；方差越小，数据的波动越小。

方差的计算公式（以样本方差为例）为：

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

在这个公式中，分母 $n-1$ 代表的是自由度——样本能自由选择的程度。当我们拥有了一组样本，假设其样本量为 10，在计算出均值后，只需 9 个数据即可将此组数据确定下来。换言之，一旦确定了均值，这组数据就只有 9 个数字是自由的，还有一个数据是被制约的，所以自由度是 $n-1$。

至于标准差，直接将方差开根号即可。

以 2005—2006 赛季 NBA 常规赛中姚明的表现为例，表 2.1 罗列了姚明在各场比赛中的得分、篮板、失误三个指标，并计算了相应的均值和方差。

表 2.1　2005－2006 赛季 NBA 常规赛姚明得分表

场次	对阵超音速			对阵快船		
	得分	篮板	失误	得分	篮板	失误
第一场	22	10	2	25	17	2
第二场	29	10	2	29	15	0
第三场	24	14	2	17	12	4
第四场	26	10	5	22	7	2
均值	**25.3**	**11.0**	**2.8**	**23.3**	**12.8**	**2.0**
方差	**8.9**	**4.0**	**2.3**	**25.6**	**18.9**	**2.7**

那么问题来了，你觉得姚明是在对阵超音速的比赛中发挥得好，还是在对阵快船的比赛中发挥得好呢？

首先来看均值。从表 2.1 中可以看到，在四场比赛中，姚明在对阵超音速的比赛中平均每场得分 25.3 分，在对阵快船的比赛中平均每场得分 23.3 分。根据这个结果，很多读者都觉得他在对阵超音速的比赛中发挥得好。

但是，篮板数和失误数也是衡量球员综合素质的很重要的指标。我们看到，在对阵超音速的比赛中，姚明的平均篮板（11 个）略低于和快船的平均篮板（12.8 个）；在对阵超音速的比赛中，每场平均失误 2.8 次，也略高于和快船队比赛中平均失误 2 次！好像问题变得复杂了。

这时，方差的作用便体现出来。从表 2.1 中可以看到，不管是在得分、篮板还是失误上，在对阵超音速的比赛中，各项指标的方差均远小于和快船比赛时各项指标的方差。一个显而易见的结论是：姚明在和超音速对战时表现得更稳定。

此时，思维比较跳跃的读者可能发现了一个更为棘手的问题：上文的例子虽然均值不同，但还是相对比较接近的（生活中，完全相等的均值几乎不会出现），可以用方差来辅助判断优劣。但如果我们对两个样本分别计算了均值和方差，发现彼此之间的均值差异很大，样本方差的差异也很大，怎么办？这就要涉及另一个概念——变异系数（$C \cdot V$）。变异系数就是用来解决均值不同（相差大）且方差也不同时如何判断孰优孰劣的问题的。

变异系数的计算方法是：

$$C\cdot V=(\text{标准差}/\text{平均值})\times 100\%$$

还是以姚明比赛的例子来说明问题，先分别计算对阵超音速和对阵快船时得分的变异系数（$C\cdot V$），如表 2.2 所示。

表 2.2　得分变异系数表

指　　标	对阵超音速	对阵快船
	得分	得分
均值	25.3	23.3
标准差	3.0	5.1
变异系数（$C\cdot V$）	11.8%	21.8%

通过计算，姚明在和超音速比赛时 $C\cdot V$ 为 11.8%，小于和快船比赛时的 21.8%，变异系数越小则说明数据越为集中，再一次证明了我们的结论。不过，当 $C\cdot V$ 指标大于 15%时，则意味着数据中可能有异常值。我们回看四场比赛的数据，在对阵快船时姚明得分仅为 17 分，远低于其他几场比赛。可见变异系数也是发现数据异常的一个好帮手。

无论是方差还是变异系数，都是数值上的比较。如果用分布图来刻画，则可以更直观地理解不同的方差（标准差）会带来怎样的视觉冲击，如图 2.3 所示。

以正态分布为例，在图中分别选取了标准差为 1、1.5、2 时的分布图，当标准差（方差）比较小时，分布曲线下的面积更加集中于均值 0 附近，分布图显得“高挑”；当标准差（方差）比较大时，数据变得更加离散，此时分布曲线的“尾部”很厚，分布图显得“矮胖”。

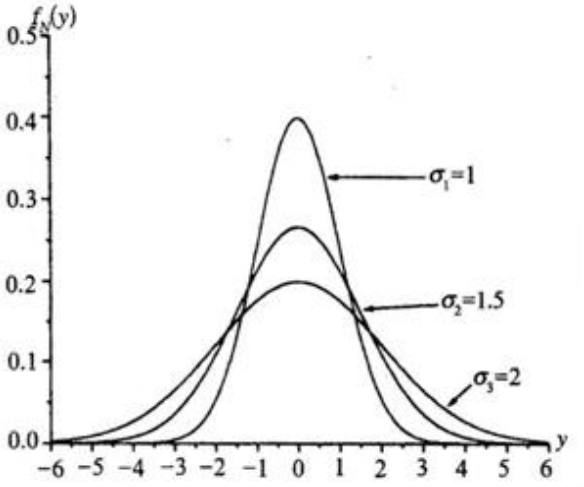


图 2.3 不同标准差对照分布图

除了标准差能刻画分布图的形状外，还有两个重要的衡量指标：峰度和偏度。

2.3 峰度&偏度——打造风度翩翩的数据分布

峰度和偏度一般会同时出现，它们的出现往往有分布图作为背景。这两个指标看似不常用，但对于数据的描述，特别是对金融数据的波动率描述有着很重要的作用。

峰度（Kurtosis）：是描述某变量所有取值分布形态陡缓程度的统计量，也是一个用于衡量离群数据离群度的指标。以正态分布与其作比较（一般正态分布的峰度为 3），如图 2.4 所示。

- Kurtosis=3：与正态分布的陡缓程度相同。
- Kurtosis>3：比正态分布的高峰更加陡峭——尖顶峰。

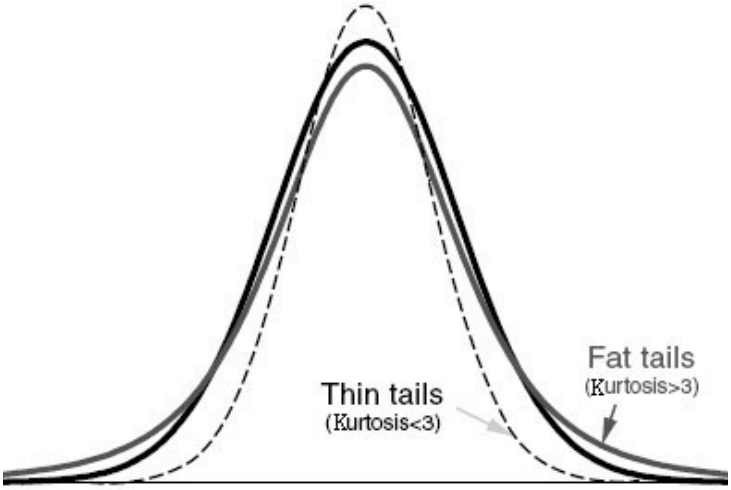


图 2.4　不同峰度对应分布图

- Kurtosis<3：比正态分布的高峰更加平缓——平顶峰。

在金融分析中，金融数据（比如收益率的分布）大多有“厚尾”倾向，同时峰度也会比较大，此时意味着这组收益率数据有较大的波动性，同时收益率也比较高。这正好印证了“高收益伴随高风险”这句话，同时也提醒大家：“股市有风险，入市需谨慎。”

偏度（Skewness）：是用来衡量数据分布是否对称的指标。

仍以正态分布为例，当数据序列呈正态分布时，它的均值、中位数和众数重合，而且在这三个数的两侧，其他所有的数据以对称的方式分布。如果数据序列的分布不对称，则均值、中位数和众数必定分处不同的位置。这时，若以均值为参照点，要么位于均值左侧的数据较多，称之为右偏；要么位于均值右侧的数据较多，称之为左偏。这和偏度又有什么关系？来看图 2.5。

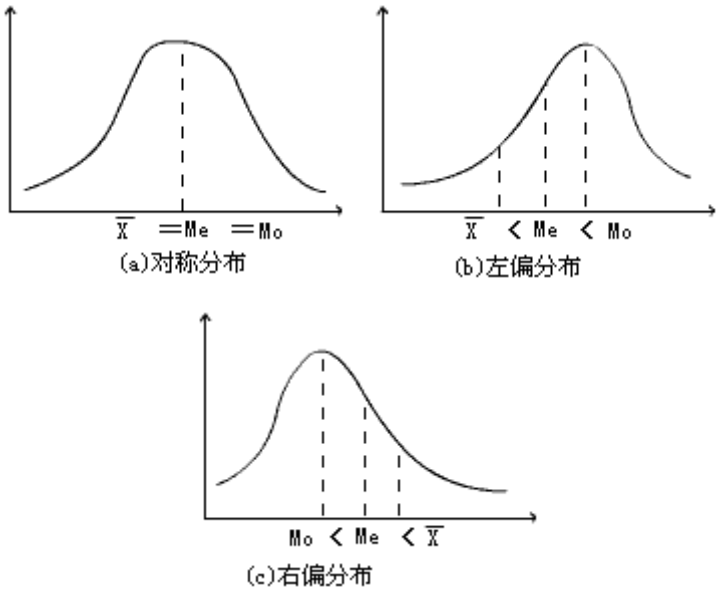


图 2.5 不同偏度对应分布图

- Skewness=0：分布形态与正态分布偏度相同。
- Skewness>0：正偏差数值较大，为正偏或右偏。
- Skewness<0：负偏差数值较大，为负偏或左偏。

不管是左偏还是右偏，只要不是对称分布，必定在分布的一侧会有较多的数据，而在另一侧数据较少。但考虑到所有数据与均值之间的离差之和应为零这一约束，当均值左侧数据较多的时候，均值的右侧必定存在数值较大的“离群”数据；同理，当均值右侧数据较多的时候，均值的左侧必定存在数值较小的“离群”数据。可以这样简单来理解：在偏度的绝对值较大的时候，最有可能的含义是“离群”数据离群的程度很高（很大或很小），这时分布曲线某侧的拖尾很长。一般地，右偏时，算术平均数>中位数>众数；左偏时则相反，众数>中位数>平均数。

如果你能够很好地运用平均数（算术平均数、中位数、众数）、方差（标准差）、峰度和偏度，再结合分位数，则能够得到数据分析的初步结果。透过数据，我们可以挖掘到更多的信息，而本章只是抛砖引玉，要想表述得更专业，则可以采用图表分析，后面的章节将详细讲述。

☆本章重点归纳

<table>
<tr><th colspan="3">指　　标</th><th colspan="2">公　　式</th><th>适用范围</th></tr>
<tr><td rowspan="6">均值</td><td rowspan="4">数值平均数</td><td>简单算术平均数</td><td colspan="2">$\frac{x_1+x_2+x_3+\cdots+x_n}{n}$</td><td rowspan="2">最常用，适合大部分情况</td></tr>
<tr><td>加权算术平均数</td><td colspan="2">$\frac{x_1f_1+x_2f_2+x_3f_3+\cdots+x_nf_n}{f_1+f_2+f_3+\cdots+f_n}$</td></tr>
<tr><td>几何平均数</td><td colspan="2">$\sqrt[n]{x_1\cdot x_2\cdot x_3\cdot\cdots\cdot x_n}$</td><td>适用于比率、增速、利率等指标计算</td></tr>
<tr><td>调和平均数</td><td colspan="2">$\frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\frac{1}{x_3}+\cdots+\frac{1}{x_n}}$</td><td>适用于平均利润率、平均成本等指标计算</td></tr>
<tr><td rowspan="2">位置平均数</td><td rowspan="2">中位数</td><td>奇数</td><td>$m_{0.5}=x_{(\frac{n+1}{2})}$</td><td rowspan="2">适用于收入、房地产价格等均值计算</td></tr>
<tr><td>众数</td><td>偶数</td><td>$m_{0.5}=\frac{x_{\frac{n}{2}}+x_{\frac{n}{2}+1}}{2}$</td></tr>
<tr><td colspan="3">方差</td><td colspan="2">$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$</td><td></td></tr>
<tr><td colspan="3">标准差</td><td colspan="2">$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$</td><td></td></tr>
<tr><td colspan="3">峰度</td><td colspan="2">$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^4f_i}{\sigma^4\sum_{i=1}^{n}f_i}$</td><td></td></tr>
<tr><td colspan="3">偏度</td><td colspan="2">$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^3f_i}{\sigma^3\sum_{i=1}^{n}f_i}$</td><td></td></tr>
</table>

第 3 章

图表的世界

如果想让你的分析报告既生动有趣，又不乏精彩的统计描述和数理推断，那么添加一些图表则是一项必不可少的专业技能。本章就来看看如何通过简单的图表给分析报告加分。

必备技能 1——频数分布表

频数分布表通常也被称作次数分布表，是工作学习中常用的统计表之一。它通常在对数据进行初步整理时用来对数据的分布进行大致归纳。

频数也好，次数也好，顾名思义，这是对数据频次刻画的统计表。如果用比较专业的用语来定义频数，则应该这样描述：频数是指某一随机事件在 n 次试验中出现的次数。各种随机事件在 n 次试验中出现的次数分布称为频数分布。

先来看一张简单的频数表，如表 3.1 所示。

表 3.1　男士鞋码频数分布表

鞋　　码	购买量（双）	占比（%）
38	10	0.9%
39	80	6.9%
40	120	10.3%
41	500	43.1%
42	300	25.9%
43	100	8.6%
44	50	4.3%
合计	1160	100.0%

此表总共有三列：第一列为标志（也可为分组标志），在本例中它按照鞋码来进行分类；第二列为各鞋码的购买数量，也就是频数；第三列则为各鞋码购买数量占总体购买量的比重。那么，这三列中哪一列才是此表最重要的考量元素？

其实，这张表最重要的就在于第一列对指标的分组分类。如何将指标恰当地分类分组决定了之后两列的频数和占比。好的分类能够让数据的分布清晰明了。在这个例子中，因为鞋码本身就有可分类性，所以一般不再做其他归类。如果想要体现数据更明显的集中趋势，也可将鞋码所在列分为：38 码以下；38～40 码；40～42 码；42～44 码；44 码以上。

对于数据的分类，可以按照列名尺度或者顺序尺度进行，通常适合对属性类指标进行分类，也可以理解为是对品质标志的分类。比如，我们需要了解近阶段医院献血人员的血型分布情况，就可以将数据按照 A 型血、B 型血、AB 型血、O 型血和 Rh 血型进行划分，然后依次统计各血型近阶段献血人次。列名尺度和顺序尺度的区别在于，列名尺度没有优劣之分，而顺序尺度则有一定的排序。比如，成绩可以按照不合格、合格、中、良、优来划分，这就是典型的顺序尺度分类。

还有一种分类方式，即按照间隔尺度或比例尺度进行，一般适合对数量进行分类。表 3.1 就是一个典型的间隔尺度。间隔尺度度量的数据一定是顺序数据，也一定是列名数据。那什么才算比例尺度呢？举个例子，统计生产线上工人每周加工的零配件个数，可将数据划分为每周加工 90～100 个、100～110 个、110～120 个、120～130 个……这样的分类就属于比例尺度。同时，在经济活动中，很多统计数据都是比例数据，如 GDP、工业总产值、主营业务收入等。那么它和间隔尺度衡量的数据有什么联系？其实很简单，间隔数据的差就是比例数据。

如果采用间隔尺度和比例尺度的分类方式，那么，画频数分布表的步骤如下。

- Step 1：求全距（用 R 表示）。

全距是原始数据中最大值与最小值之差，数学表达式为：$R=\max\{x_i\}-\min\{x_i\}$。式中，$R$ 是全距，$\max\{x_i\}$为这组数据中的最大值，$\min\{x_i\}$为这组数据中的最小值。

- Step 2：定组数（用 K 表示）。

组数就是对这组数据分组的个数。一般来说，制作一张频数分布表，通常把数据分为 10 组左右。

- Step 3：定组距（用 i 表示）。

组距是指每个组内包含的间距，即组距（i）表示每个小组的组上限（组的终点值）与组下限（组的起点值）之间的距离。全距、组数和组距的关系为：

$$i=\frac{R}{K}$$

当全距固定之后，组数越多，组距越小；组数越少，组距越大。

- Step 4：列组限。

组限是每一组在数尺上的起始点和终止点，即上下限。这里需要特别注意，在进行分组的时候，一定要遵循一个原则：不重不漏。“不重”说的是任何一个数据都只能被分在一组中，不能同时出现在两组中；“不漏”说的是每一个数值必须被归类，不能“掉队”。

- Step 5：求出组中值（用 m_0 表示）。

组中值是各组的中点值。组中值等于该组的组限右端点与左端点的值的平均数。

- Step 6：归组登记频数（用 f 表示）。

至此，频数分布表就绘制完成了。接下来举一个简单的例子。

假设有一个零部件加工车间，统计了它两年内每月加工的零部件个数，经过排序整理后，具体数据如表 3.2 所示。

表 3.2　月度零部件加工数据

20 277	20 886	23 007	27 289	35 232	35 716	36 406	43 195	47 487	50 485	58 158	58 645
60 350	62 385	66 285	66 532	72 673	74 155	77 781	79 290	87 224	88 173	93 293	97 413

从表 3.2 中可以看到，月加工零部件最少的月加工了 20 277 个，最多的月加工了 97 413 个，则可以通过计算得到：R=97 413–20 277= 77 136。

我们来看看这组数据：总量为 24 个数据，全距为 77 136。综合

考虑，我们把组数固定为 8 组，这样通过公式就可以得到组距为 9642。为了方便划分，选取 10 000 作为组距。

现在有了这几个关键的数据，就可以进行上下限的划分，频数分布表即可确定，如表 3.3 所示。

表 3.3 月度零部件加工数频数分布表

分　组	*f*	百分比	累积百分比
30 000 以下	4	17%	17%
30 000～40 000	3	13%	29%
40 000～50 000	2	8%	38%
50 000～60 000	3	13%	50%
60 000～70 000	4	17%	67%
70 000～80 000	4	17%	83%
80 000～90 000	2	8%	92%
90 000 以上	2	8%	100%
合计	24	100%	

在上述分组中，有两组很特别，即 30 000 以下和 90 000 以上，这两组都没有闭合的组限，称之为开口组。在组距上，本例采用的是等距分组，而在生活和工作中，有时也会选择采用不等距分组，比如，当数据分布不均或者有特殊规定时。

必备技能 2——频数分布图

常用的频数分布图一般为频数直方图。频数直方图是建立在频数分布表的基础上的。在绘制完表格后，可以通过 Excel、SPSS 等数据分析软件来绘制频数直方图。

以零部件加工频数分布表和 SPSS 统计分析软件来做一个示范。首先将数据导入 SPSS 统计分析软件，依次单击图形→旧对话框→直

方图，在弹出的窗口中将需要进行绘图的变量选入变量框，单击确定后即可得到频数分布直方图，如图 3.1 所示。

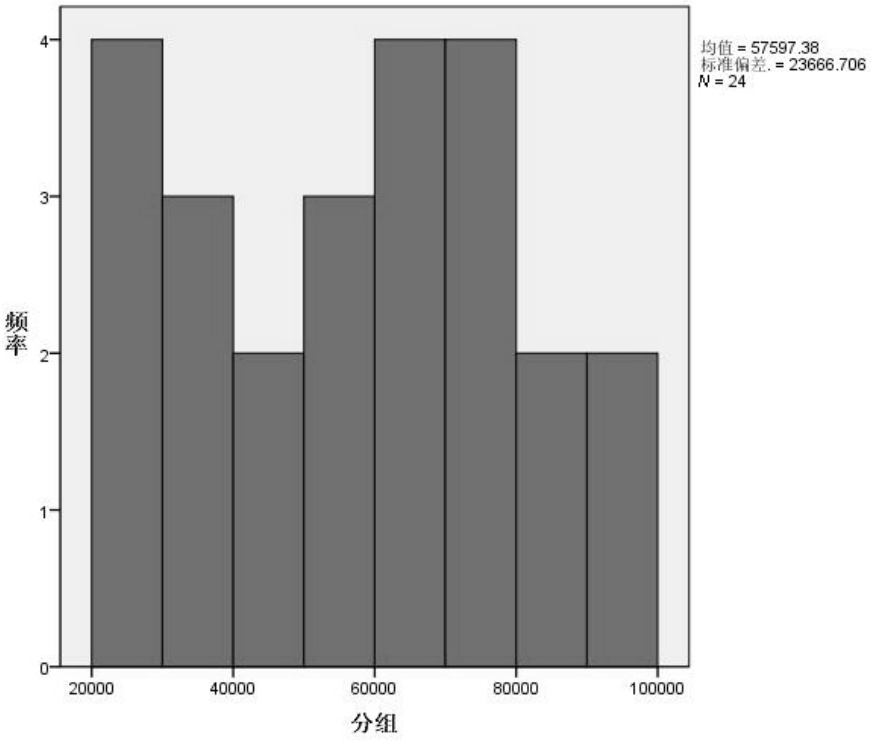


图 3.1　月度零部件加工频数分布直方图

直方图的横轴代表分析变量数据的频数区间，纵轴代表每个区间的频数。从图形中可以直观地看出各个区间的频率分布情况。

直方图和条形图形状类似，但二者的区别如下：

- 条形图用条形的长度表示各类别频数的多少，用宽度来表示类别，是固定的；直方图用矩形的高度表示每一组的频数或频率，宽度则表示各组的组距，因此其高度与宽度均有意义。

- 由于分组数据具有连续性，直方图的各矩形通常是连续排列的，而条形图则是分开排列的。
- 条形图主要用于展示分类数据，而直方图主要用于展示数值型数据。

如果用同样的数据画一张条形图，则其形状如图 3.2 所示，无法体现分组的概念。

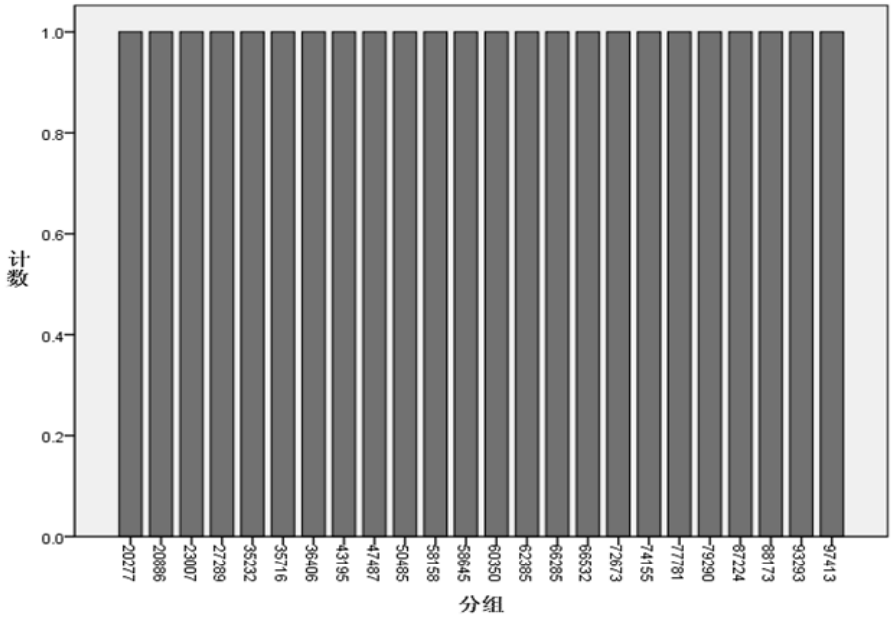


图 3.2　月度零部件加工频数分布条形图

换一组数据，使用鞋码购置数量来绘制条形图，则其形状如图 3.3 所示，图形较为直观。

必备技能 3——茎叶图

茎叶图又称枝叶图，它的绘制思路很形象，也很具体化，概括起来就是将数据按位数进行比较。将数的大小基本不变或变化不大的位数作为一个主干（茎），将变化大的位数作为分枝（叶），列在茎的后面，这样就可以清楚地看到每个主干后面有几个数、每个数具体是多少。

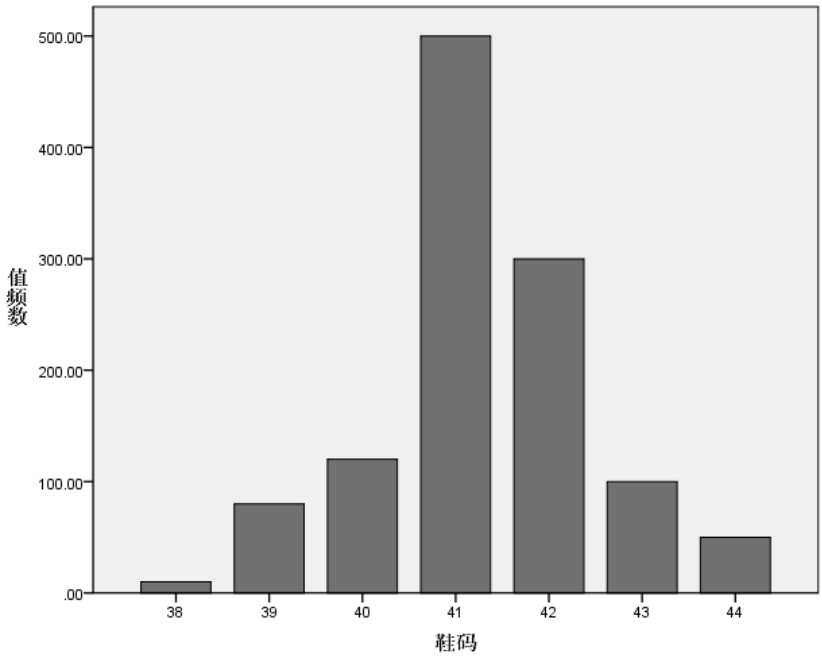


图 3.3　男士鞋码频数分布条形图

简单来说，茎叶图有三列数：左边的一列数一般是频数统计；中间的一列表示茎，也就是变化不大的位数；右边的一列是数组中的变化位，它会把数组中每个变化的数一一罗列出来。如此一来，整幅图就像一条枝上抽出的叶子一样。

如果将茎叶图的茎和叶逆时针旋转 90°，那么它实际上就是一张直方图，可以从中统计出次数，计算出各数据段的频率或百分比，进一步可以看出分布是否与正态分布或单峰偏态分布逼近。当然，茎叶图和直方图也有显著的区别，茎叶图可以看到数据的细节；而直方图虽然直观、简单，但却丢失了数据的原始信息。

另外，茎叶图还有以下两个特点：

- 茎叶图中的数据可以随时记录、随时添加，方便记录与表示。
- 茎叶图只便于表示个位之前相差不大的两组数据。

为了具体展现第二个特点，我们用两组数据来进行对比。

第一组数据：**20　20　23　27**　35　35　36　43　47　50　58　58

60　62　66　66　72　74　77　79　87　88　93　97

通过 SPSS 统计分析软件得到的茎叶图如图 3.4 所示。

Stem-and-Leaf Plot

Frequency	Stem &	Leaf
4.00	**2 .**	**0037**
3.00	3 .	556
2.00	4 .	37
3.00	5 .	088
4.00	6 .	0266
4.00	7 .	2479
2.00	8 .	78
2.00	9 .	37

Stem width: 10

Each leaf: 1 case(s)

图 3.4　数据茎叶图

以第一行为例，在这组数据中，十位数以 2 开头的共有 4 个，分别是 20、20、23 和 27，所以频数为 4；同时 Stem（茎）部分显示为 2，代表 20；Leaf（叶子）部分分别为 0、0、3 和 7，这也就是这 4 个数字的个位数——是作为区分以十位数 2 开头的 4 个数字的变化部分。

图下方的 Stem width 表明了茎部分的数位；Each leaf：1case 表示叶子部分的每一个数字出现一次。在两位数的数据中，我们发现，茎叶图不仅可以展示数据的频数分布情况，还可以知道各分类数据的具体分布情况。不过，当数据的尾数扩大至三位以上时，情况就大不相同了。我们来看通过第二组数据绘制的茎叶图，如图 3.5 所示。

Stem-and-Leaf Plot

Frequency	Stem & Leaf
4.00	2 . 0037
3.00	3 . 556
2.00	4 . 37
3.00	5 . 088
4.00	6 . 0266
4.00	7 . 2479
2.00	8 . 78
2.00	9 . 37

Stem width: 10000

Each leaf: 1 case(s)

图 3.5　月度零部件加工数茎叶图

神奇的一幕发生了，图 3.5 竟然和图 3.4 几乎一样，而这组数据正是之前制作频数分布表时所使用的零部件每月加工数的数据。这张图的茎部数据是需要乘以 10 000 的，也就是说茎部的 2 表示的是

20 000，枝叶部分的数据表示千位的数据。

为什么会这样？回到我们之前对茎叶图的特点描述，它的茎部使用的是数值相对固定的位数，枝叶部分使用的是变化较大的部分数据，以此来对数据进行分类。结合到零部件案例，因为有 4 个月都是月加工 20 000 多个零部件，对于这 4 个月，万位数是固定不变的，区别是从千位数开始的，这也就造成了使用两组完全不同的数据，却画出了几乎一样的茎叶图，此时茎叶图展示数据细节的作用就不那么大了。

必备技能 4——箱线图

箱线图，又称箱形图，其形状如图 3.6 所示。

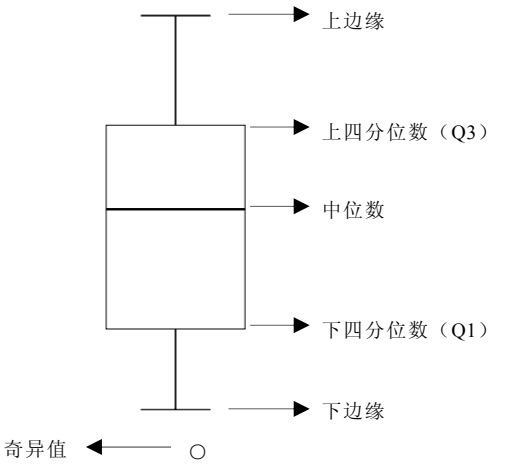


图 3.6　箱线图图示

箱线图的最大特点是它能通过几根线和一个箱体来描述数据的分布（是对称分布还是左偏、右偏分布）和集中度情况，并且能够通

过此图来判断分析是否存在异常值。绘制箱线图的过程其实也就是计算 5 个数的过程。

箱线图的主要绘制步骤如下。

- Step1：画数轴。
- Step2：画矩形盒。

两端边的位置分别对应数据的上、下四分位数（Q1 和 Q3）。在矩形盒内部的中位数位置画一条线段为中位线。

- Step3：在 Q3＋1.5IQR（四分位距）和 Q1−1.5IQR 处画两条与中位线一样的线段，这两条线段为异常值截断点，称其为内限；在 Q3+3IQR 和 Q1−3IQR 处画两条线段，称其为外限。处于内限以外位置的点表示的数据都是异常值，其中在内限与外限之间的异常值为温和的异常值（用“〇”标出），在外限以外的为极端的异常值（用“＊”标出）。
- Step4：从矩形盒两端边向外各画一条线段直到不是异常值的最远点，表示该组数据正常值的分布区间点，即该组数据正常值的分布区间。

仍以零部件加工为例，假设从第三年开始，该车间突然接到一笔大订单，于是车间开始满负荷生产。最终，该月车间加工的零部件个数达到 198 772 个。我们对这一新增的数据，也就是共 25 个数据使用 SPSS 统计分析软件绘制箱线图，如图 3.7 所示。

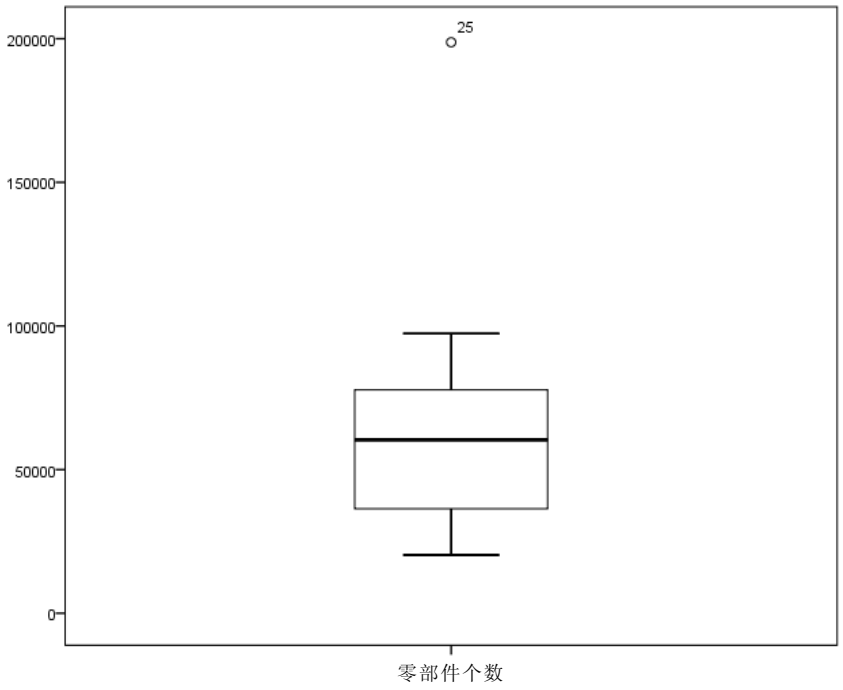


图 3.7 箱线图

这张箱线图最突出的部分就是顶端的那个小圆点，这就是一个异常值，而且从图形上来看，整个箱体也不太对称。在这个小圆点上还标着数字 25，这正是我们加入的第 25 个数据。所以，通过箱线图不仅可以知道数据中是否有异常值，还能快速地知道哪一（几）个是异常值、偏离的程度是多少。对箱线图的作用进行归纳，主要有以下三点。

- 有效识别数据是否存在异常值。
- 直观判断数据是否呈现偏态。
- 快速比较几组数据的形状分布。

必备技能 5——散点图

散点图和前面讲述的几张图的最大区别是它涉及对两个变量之间相互关系的描绘。在回归分析中，散点图运用最多，同时也往往是进行数据分析的第一步。可以通过散点图来大致判断变量是否呈现线性相关，观察变量间大致的变化趋势，还可以为选用什么样的模型提供参考。

散点图并不难画，只需将两个变量分别作为 *X* 和 *Y* 坐标轴即可轻松画出。为了让读者有个直观的了解，下面通过一个小案例来进行解析。

选取身高和体重两个指标各 20 个数据组成数据集，如表 3.4 所示。

表 3.4　身高体重数据表

身高（cm）	155	156	176	166	167	154	177	146	167	166	176	144	156	166	155	153	156	167	158	160
体重（kg）	55	56	70	56	70	44	66	45	70	68	78	42	55	60	60	54	53	60	49	45

有了数据，我们可以通过 SPSS 来绘制散点图。依次选择图形→旧对话框→散点点状→简单分布，将身高作为 *Y* 轴，体重作为 *X* 轴，单击确定后即可得到散点图，如图 3.8 所示。

为了更好地理解散点图对于变量间关系的描绘，图中特地加了一条辅助线。结合这条辅助线，可以看到，散点图中身高和体重之间确实有着一些线性关系，虽然与 *Y*=*X*+100 这条辅助线之间还有一些出入，但整体趋势是一致的，图形表现为带状，呈现向右上方倾斜的走势。

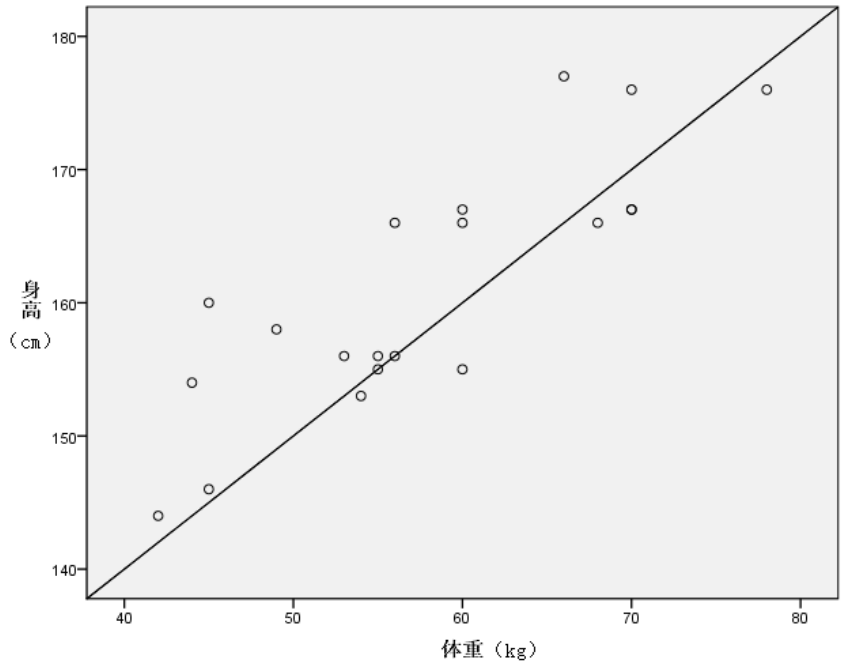


图 3.8 身高体重散点图

我们说散点图可以描绘两个变量的相关情况，因为我们的绘图是通过两个坐标轴来制定的，但这并不代表只有两个变量才能绘制散点图，多变量也可以绘制散点图，因而出现了三维散点图和散点矩阵图。比如，在上述数据中再增加一组年龄变量，现在所组成的数据集如表 3.5 所示。

表 3.5 身高、体重及年龄数据表

身高（cm）	155	156	176	166	167	154	177	146	167	166	176	144	156	166	155	153	156	167	158	160
体重（kg）	55	56	70	56	70	44	66	45	70	68	78	42	55	60	60	54	53	60	49	45
年龄（岁）	30	25	26	36	64	19	36	21	65	60	47	17	24	46	45	28	32	56	22	18

在选择散点图时依次选择图形→旧对话框→散点点状→矩阵分布，然后将三个变量全部作为矩阵变量，就可以得到散点矩阵图，如图 3.9 所示。

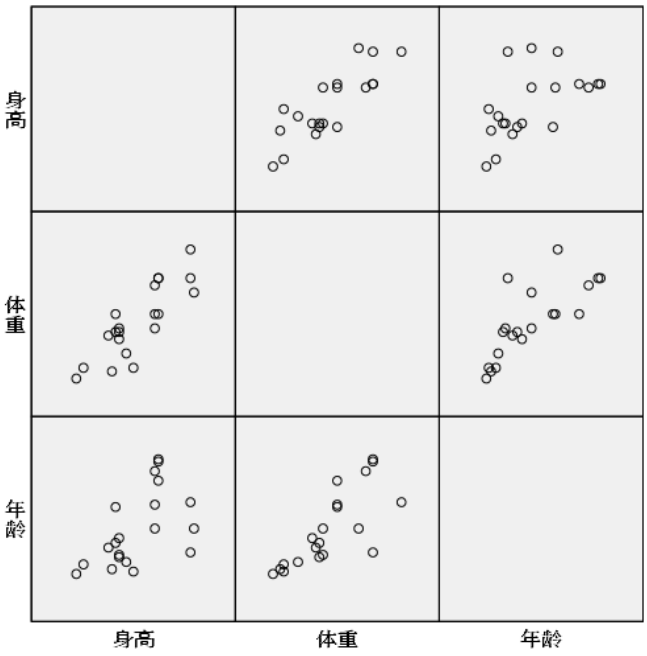


图 3.9　身高、体重、年龄散点矩阵图

这幅图把三个变量进行两两组合，绘制在一个矩阵里，从而能够方便地看出各个变量中哪些变量之间的相关程度比较高。

比如，可以从图中看到身高和体重之间的线性关系明显强于身高和年龄之间的线性关系。在增加了年龄变量之后，可以看出，体重和年龄之间也存在一定的线性关系，至于这个关系到底如何，它与体重和身高之间的关系相比孰强孰弱，就得交给回归分析和相关分析来进行量化统计了。

除了数据变量原本客观存在的相关程度外，还有一个因素也会影响散点图的呈现质量，那就是数据量。通常情况下，数据量越多，散点图的表现优势越显著。设想一下，如果两个变量之间的关系可以完全用一个线性函数来表示，那么它画出来的应该是一条直线。如果只用 5 个观测值来画，就会是 5 个点（虽然它们可以连成一条直线）；但假如用 1000 个观测值来画，就可能会是一条较为光滑的直线。

至此，5 个数据分析必备技能已经介绍完毕，不过只学会这些技能还不够，能否正确使用才是最重要的。下面罗列一下在图表绘制中经常忽略的小细节。

细节 1：数据范围选择正确吗？

在分析数据时，并不一定需要选择所有的数据，特别是当我们分析的是经济类数据时，往往会选择某一段时间的数据。此时如果不对数据范围进行合适的筛选，就有可能对他人造成误导。举一个例子，如图 3-10 所示。

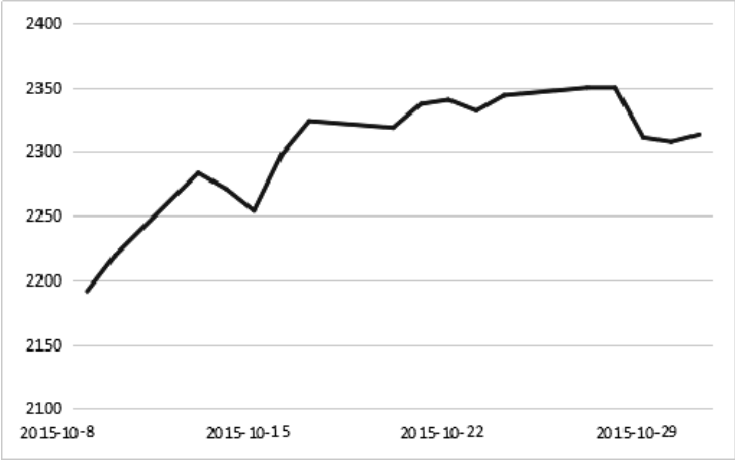


图 3.10　2015 年 10 月上证 50 收盘价折线图

从图 3.10 中可以看出，整个指数走势是向上的。这是一个指数的走势，如果它向上挺进，则是否意味着股市也在向牛市发展？我们将数据的范围扩大几个月再来看看，如图 3.11 所示。

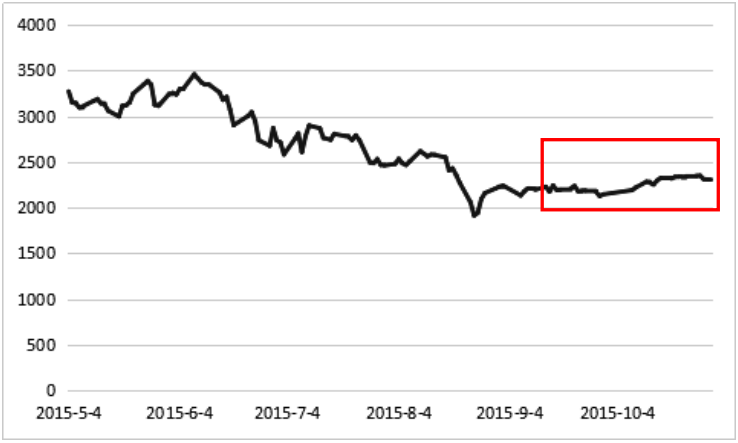


图 3.11　2015 年 5—10 月上证 50 收盘价折线图

这张图的时间跨度从 2015 年 5 月至 2015 年 10 月，历时半年。在这张图上，在图 3.10 中明显的上升趋势缓和多了。再仔细看一下 2015 年 5—6 月的上证 50 收盘价，对比红框内的折线图，可以清楚地发现，这个缓慢上升并非意味着股指的大涨，反而更像大跌后的反弹。

可见，如果选择的数据范围缩小，虽然可以更加细化地看出这个范围内数据的变动趋势，但是如果放到更广阔的范围中，得出的数据分析结论可能会是完全不同的，需根据实际情况来界定。

细节 2：*Y* 轴起点在哪里？

很多时候会将 *Y* 轴的起始点定为 0，不过在实际作图中，*Y* 轴起始点并非是 0，这对图形又会有什么影响呢？

在图 3.10 中，*Y* 轴的起始点在 2100；而在图 3.11 中，*Y* 轴的起始点变为 0。其中一个原因是在 2015 年 8 月指数跌破了 2000 关口，这必然造成 *Y* 轴的刻度会随着改变。

如果更改一下刻度比例，将 *Y* 轴的取值区间更改为[1500,3500]，此时的图形如图 3.12 所示。

图 3.12　坐标更改折线图

这样一看，折线图的走势更为陡峭，而且 8 月份的那次杀跌在图中表现得更为突出。可见，更改了坐标轴，虽然对数据来说没有变动，但是对图形的直观感受却是有显著影响的。

细节 3：比例真的适合你的数据？

我们所说的比例其实是图表的比例（并非上文的刻度比例）。什么是图形比例？简单来说就是图表的形状。先来看这样一组图表，如图 3.13 所示。

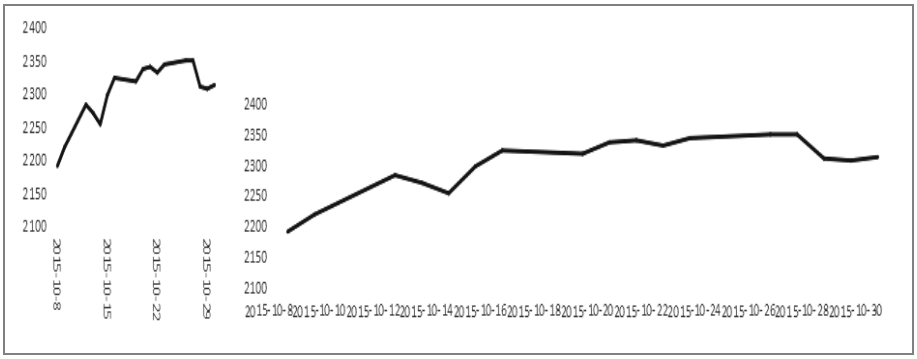


图 3.13　比例更改对比图

这两张图其实是同一张折线图，通过对比不难看出，左侧图的数据波动较为明显，而右侧图的数据波动则显得比较平坦。仅仅是图片比例的变动，也能引起截然不同的感官刺激。

在实际工作中，还有一些绘图注意事项，如下：

- 需要根据数据的性质和分析目的来选用适当的图表。
- 图表的标题应说明数据的内容，一般位于图的下方。
- 图表的横、纵轴应注明标目及对应单位，尺度应等距或具有规律性，一般自左而右、自上而下、由小到大。
- 为使图形美观并便于比较，统计图的长宽比例一般为 7∶5，有时为了说明问题也可加以变动。
- 比较、说明不同事物时，可用不同颜色或线条表示，并常附图例说明，但不宜过多。

☆本章重点归纳

- 条形图：表示各个项目之间的对比。可细分为如下几类。
  - 簇状条形图：用于比较类别间的值。它也可以以三维效果显示。一般水平方向表示类别，垂直方向表示各类别的值，从而关注值的对比情况。
  - 堆积条形图：用于显示各个项目与整体之间的关系。
  - 百分比堆积条形图：以百分比形式比较各类别的值在总和中的分布情况。
- 折线图：按照相同间隔显示数据的趋势。可细分为以下两类。
  - 堆叠折线图：用于显示各个值的分布随时间或类别的变化趋势。
  - 百分比堆叠折线图：以百分比方式显示各个值的分布随时间或类别的变化趋势。
- 饼图：用于显示组成数据系列 的项目在项目总和中所占的比例。饼图通常只显示一个数据系列，当希望强调数据中的某个重要元素时可以采用饼图。可细分为以下两类。
  - 分离型饼图：用于显示各个值在总和中的分布情况，同时强调各个值的重要性。
  - 复合饼图：这是一种将用户定义的值提取出来并显示在另一张饼图中的饼图。例如，若为了看清楚细小的扇区，就可以将它们组合成一个项目，然后在主图表旁的小型饼图

或条形图中将该项目的各个成员分别显示出来。

- XY 散点图：显示若干数据系列中各数值之间的关系，或者将两组数绘制为 *XY* 坐标的一个系列。散点图通常用于科学数据。
- 面积图：强调大小随时间发生的变化。
- 圆环图：像饼图一样，圆环图显示部分和整体之间的关系，但是它可以包含多个数据系列。
- 雷达图：用于显示数值相对于中心点的变化情况。显示时可以为每个数据点显示标记。

以上即为各类图形特点汇总。

第 4 章

当小“正太”遇上“大叔”——正态分布篇

在统计学领域，小“正太”指的是正态分布；而“大叔”则有两位，分别是大数定律和中心极限定理。

之所以要花一个章节的篇幅来讲正态分布，原因就在于它是统计学的基础，特别是在推断统计发展中，它是一块基石。可以毫不夸张地说，统计学中的理论知识（包括前面提到的抽样调查技术等）大多数是建立在数据服从正态分布的前提条件下的。

在正式介绍“正太”之前，先来了解一下它的家族——分布。

分布，从字面上解释，其实就是在一定区域范围内的散布情况。在统计学中，分布一般指的是概率分布。

那什么是概率呢？先来认识一个名词：随机变量。

随机变量就是某个变量的取值是随机的。如果放置在生活中，那么这个变量可以通俗地描述为今天是下雨天还是晴天；专业一点则可以这样表述：做一次试验，会出现多种可能的结果，每一种可能的结果都可以用一个数来表示，把这些数作为变量 x 的取值范围，则试验结果可用变量 x 来表示，同时这个变量也称为随机变量。

随机变量是一个很大的变量集合体，它还可以细分为离散型随机变量和连续型随机变量。如果表示试验结果的变量 x，其可能取值为有限多个，且都有确定的概率，则称 x 为离散型随机变量；如果表示试验结果的变量 x，其可能取值为某范围内的任何数值，且 x 在其取值范围内的任一区间取值时，其概率是确定的，则称 x 为连续型随机变量。

了解了随机变量的定义，就可以进一步介绍“概率”这个概念了。概率是对随机事件发生的可能性的度量，一般以一个 0～1 的实数表示一个事件发生的可能性大小。该值越接近 1，该事件越可能发生；越接近 0，则该事件越不可能发生。这里怎么是随机事件？其实，随机变量和随机事件之间的关系可以用一句话来概括：随机变量的取值代表了随机事件。所以刻画随机事件的概率其实也可转换为刻画随机变量的概率。正因如此，诞生了随机变量概率分布函数，用来描述随机变量概率在取值范围内的散布情况。

有了上述的知识铺垫，我们只需知道小“正太”——正态分布是一个很重要的连续型随机变量的概率分布即可。

4.1 小“正太”的基本情况

姓名：正态分布。

昵称：高斯分布。

出生时间：十八、十九世纪。

要说小“正太”的发现者是何许人也，这在学术界曾有过激烈的争论。早在 1738 年，棣莫弗便发表了二项式分布在特殊情况下趋近于正态分布的结果，但并未引起世人注意。拉普拉斯于 1778 年发表了误差的频率与其平方的指数呈正比的结论，1782 年又计算了正态分布的归一化常数，1810 年又发表了中心极限定理的原始版本，现在被称为棣莫弗-拉普拉斯定理。而高斯则于 1809 年发表了最小二乘法、最大似然估计及正态分布，并且严格证明了误差服从正态分布，不过他宣称自己早在 1795 年就发现了这些结果。

其实，除了棣莫弗、拉普拉斯与高斯三人，正态分布的历史还涉及众多人。而且一个理论的形成不是独立的，往往能引申出一连串相应的理论或推论，如在统计学中也有同样重要地位的最小二乘法。勒让德 1805 年就独立发表了最小二乘法。更令人惊讶的是，其实美国数学家 Robert Adrain 与高斯在同一时间得到了正态分布的结论，只可惜当时未被世人知晓，直到 1871 年另一个美国人、气象学家 Abbe 发现了这一结论。其实，是高斯拓展了最小二乘法，把正态分布和最小二乘法联系在一起，并在统计误差分析中确立了正态分布的定位。

4.2 小“正太”的性格和优点——正态分布的定义和特征

首先来介绍一下正态分布到底是一个怎样的分布，数据符合哪些条件才能称为正态分布。

在数学定义中，若随机变量X服从一个数学期望为μ、标准方差为σ的高斯分布，记为$X\sim N(\mu,\sigma^2)$，其概率密度函数为$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$，这就是一个正态分布的概率密度函数。

为什么称这个分布为正态分布？其实它是高斯函数的一个展现，如果引用它的英文名可能更容易理解：Normal Distribution。可以看到，在定义中有用对均值（用统计术语来说是期望值）和标准差这两个参数来描述正态分布，事实上也正是由这两个参数来决定分布图形状的。在日常生活、工作中，凡事都会有一个标准，在一个群体中，合乎标准的占多数，偏离标准的占少数，偏离标准越远，出现的数量就越少，这叫作事物属性的正常分布状态，简称正态。至此，小“正太”终于得到正名了。

再来说说均值μ和标准差σ，它们分别决定了分布曲线在坐标轴上的位置和变动幅度。正态分布的图形可以大致这样描述：一组遵从正态分布的随机变量，它的概率规律为取μ附近的数值概率比较大，而取离μ越远的数值概率比较小。如果σ越小，那么数据分布越集中在μ附近；σ越大，数据分布越为分散。

正态分布的密度函数的特点是：关于μ对称，在μ处达到最大值，在正（负）无穷远处取值为0，在$\mu\pm\sigma$处有拐点。它的形状是中间高两边低，图像是一条位于X轴上方的曲线。

如果把正态分布的概率密度分布图绘制出来，那么它看上去像一口钟，所以正态分布曲线又称钟形曲线，如图4.1所示。如果μ=0，σ=1，则称之为标准正态分布。为什么称之为标准？因为它正好是关于Y轴对称的。

图 4.1 中那条 F 曲线就是标准正态分布曲线，而其他几条颜色的曲线则是 μ 和 σ 取不同值时的分布曲线。

正态分布的主要特征如下。

- 集中性：正态曲线的高峰位于正中央，即均值所在的位置。

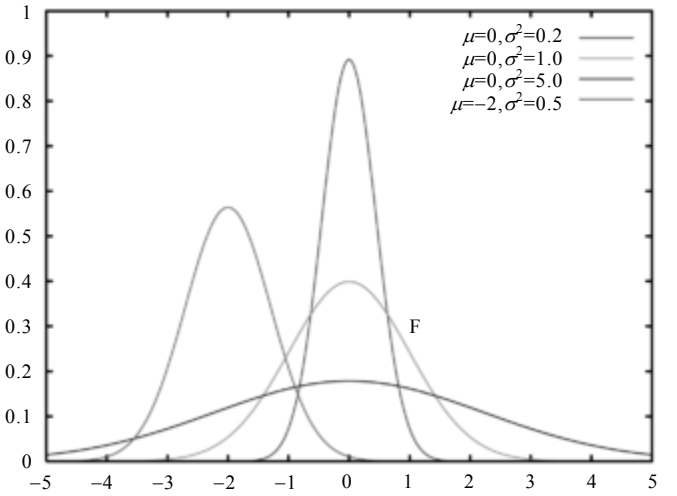


图 4.1 正态分布曲线

- 对称性：正态曲线以均数为中心，左右对称，曲线两端永远不与横轴相交。

- 均匀变动性：正态曲线由均数所在处开始，分别向左右两侧逐渐均匀下降。

- 正态分布有两个参数，即均值 μ 和标准差 σ。均值 μ 决定正态曲线的中心位置；标准差 σ 决定正态曲线的陡峭或扁平程度。σ 越小，曲线越陡峭；σ 越大，曲线越扁平。

- U 变换：为了便于描述和应用，常将正态变量作数据转换。

- 在正态曲线下方和 X 轴上方范围内区域面积为 1。

其中，U 变换是正态分布的一个重要特性，它可将一般正态分布转化为标准正态分布。这个转化是如何构造的呢？首先需要知道的是，一个正态分布经过线性变换后依旧服从正态分布。假如一组数据 X 服从的是 $N(\mu,\sigma^2)$，那么可以构建一个变量 U，使之形式为 $U=\frac{X-\mu}{\sigma}$，这时候这个变量 U 服从的就是 $N(0,1)$ 的标准正态分布。

4.3 小“正太”的可爱之处——正态分布的作用

下面介绍一下正态分布的作用。

1. 可用于质量控制

从我们刚刚罗列的最后一条性质出发，如果再拓展引申一下，便可与质量管理体系中非常有名的两个准则——“3σ 准则”和“6σ 准则”结合起来。

统计学和质量管理体系又是怎么结合起来的呢？这就涉及它们之间的一个桥梁专业：质量控制与管理。在质量控制与管理学中有大量的理论和统计学相容。

细心的读者可能会在这两个准则中发现一个熟悉的符号——标准差 σ。这两个准则其实就是建立在生活中大量事物的发生概率服从或近似服从正态分布的基础上的，如考试成绩、身高、体重等，都服从正态分布。那问题来了：服从正态分布与质量管理有什么关系呢？先来看一张图，如图 4.2 所示。

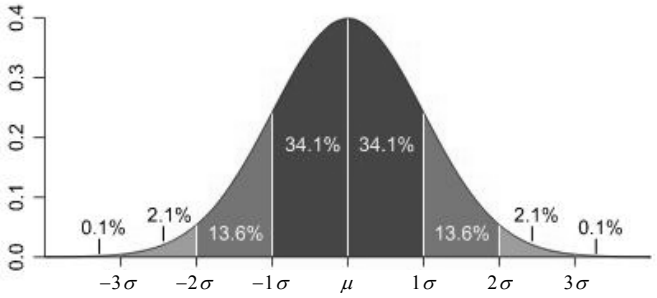


图 4.2 数值分布图

从图 4.2 中可以归纳出以下三点：

- 数值分布在$(\mu-\sigma,\mu+\sigma)$区间的概率为 0.6827。
- 数值分布在$(\mu-2\sigma,\mu+2\sigma)$区间的概率为 0.9545。
- 数值分布在$(\mu-3\sigma,\mu+3\sigma)$区间的概率为 0.9973。

也就是说，可以认为，如果一组数据服从正态分布，那么它的取值几乎全部集中在$(\mu-3\sigma,\mu+3\sigma)$区间内，超出这个范围的可能性仅占不到 0.3%。如果超出这个范围，则称为“小概率事件”。“小概率事件”通常是指发生的概率小于 5%的事件，也就是说，一般情况下，在一次试验中，该事件是几乎不可能发生的。

关于“小概率事件”，要有两个方面的认识：第一，这里的“几乎不可能发生”是针对“一次试验”来说的，如果试验次数多了，该事件是很可能发生的；第二，在运用“小概率事件几乎不可能发生”的原理进行推断时，也有 5%的犯错概率。而在“3σ 准则”中，这样的犯错概率被控制在 3%以下，使得它成为一个更为可靠的质量检验标准。

那“6σ 准则”呢？其要求更高，数值分布在(μ–6σ,μ+6σ)区间的概率为 0.999999998，近乎为 1。假如我们对某种产品的合格率检验采取“6σ 准则”（产品合格率一般也服从正态分布），那么这个检验是相当严格的。

目前有很多工业生产领域的企业采用了“6σ 准则”来作为品质控制标准。

2．可用于均值等参数的估计

使用样本参数来估算总体参数的做法，在统计学中有个专业术语：点估计。

在统计学中还有一个更为常用的估计方法——区间估计法，它是在点估计的基础上加上一个区间范围，这样就可以扩大总体参数落在估计范围内的概率。而在描画参数区间这件事上，正态分布又起到了重要的作用。

首先来了解一下进行区间估计之前需要哪些资料。区间估计所需的统计资料不多，以均值的区间估计为例，只要知道样本容量、样本均值、样本方差即可。当然，如果知道总体和样本服从的分布更佳，如果不知道，也可以通过样本容量来选择相应的估计方法。下面通过一个简单的例子进行介绍。

一家灯泡制造商每天大概能生产 10 000 只灯泡，为了估算灯泡的使用时间是否符合标准，抽取 25 只来进行检验。按照原来的技术规定，灯泡使用时间的标准大约为 2000 小时，而且总体的标准差是 10，服从正态分布。经过检验发现，这批样本的平均使用时间为 2005 小时，若采用 95%的置信水平，那么该如何估算这 10 000 只灯泡的

置信区间呢？

我们来梳理一下该例子中的一些信息。

首先说两个重要的概念：置信区间和置信水平。

置信区间就是在区间估计中，由样本统计量所构造的总体参数所在的区间。比如以 25 只样本灯泡的使用时间来估计 10 000 只灯泡的平均使用时间，求得的这个区间就是一个置信区间。

而置信水平是指将构造置信区间的步骤重复多次，置信区间中包含总体参数真值的次数所占的比率。举个例子，如果我们有足够的时间和精力做 100 次抽样，获得 100 个样本参数，那么可以计算出 100 个置信区间。如果设定置信水平为 95%，也就是说有 95%的区间包含总体参数的真值，那么还有 5%的真值可能没有包含在内。

回到我们的例子，抛开这两个概念，则可以大致整理为：总体样本为 10 000 只灯泡，样本容量 n=25，样本均值为 2005，置信水平为 95%。

那么，由这些数据如何计算置信水平呢？我们给出置信区间的计算公式：

$$(\bar{x}-z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\bar{x}+z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}})$$

其中，α 是事先确定的一个概率值，也称作显著性水平，它是总体参数的真值不落在置信区间的概率。很容易得出，$1-\alpha$ 就是我们所需的置信水平。本例中置信水平是 95%，那么 α=0.05。

$z_{\frac{\alpha}{2}}$是标准正态分布中侧面积为$\frac{\alpha}{2}$时的 Z 值，在本例中所求的就是侧面积为 0.025 的 Z 值，即 1.96，该值可以通过查标准正态分布表来获得。

$z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$是估计的允许误差，又称估计误差或误差范围。

将条件中的数据代入，即可得到$(2005-1.96\times\frac{10}{\sqrt{25}},2005+1.96\times\frac{10}{\sqrt{25}})$，进一步计算即可得到总体的平均使用时间的置信区间为(2001.08,2008.92)。

相比样本平均使用时间 2005 小时，我们给出的总体均值的区间更可靠。通过标准正态分布表得知，我们将自己的估算精度提高了。小“正太”是不是很有用？

3．可以参与制定参考值范围

对于服从正态分布指标，我们还可以用它来制定一些参考值的范围，比如我们经常在各种医学检查中看到检查项目的正常参考值。它们是怎么利用正态分布来制定的呢？

医学参考值范围，我们也通常称之为医学正常值范围，是指所谓“正常人”的解剖、生理、生化等指标的波动范围。制定正常值范围时需要如下几步：

首先要确定一批样本含量足够大的“正常人”。所谓“正常人”，不是指一点小病都没有的“健康人”，而是指排除了影响所研究指标的疾病和有关因素的人群。

其次要设定参考值范围制定所需的样本数。通常情况下，此类样本数不低于 100。

再次要对选定的正常人进行统一而准确的测定。这里的“统一”有两层含义：第一，对测定的方法、仪器、试剂和操作方法的精确度要统一；第二，要尽量与应用医学参考值范围时的实际情况相统一。这时就要考虑是否要按照性别、年龄等因素进行分组测定。对于一些检测来说，这些因素往往会引起组间明显的差异，当然最后的参考值也需要做相应的微调。

最后要根据研究目的和使用要求选定适当的百分界值。不仅如此，还需要进一步根据指标的实际用途确定单侧或双侧界值。另外，还要根据资料的分布特点，选用恰当的计算方法。常用的计算方法有两种。

（1）正态分布法：适用于（近似）正态分布或对数正太分布的资料。

（2）百分位数法：常用于偏态分布资料及资料中一端或两端无确切数值的资料。

4．众多统计方法的理论基础

尽管正态分布一直被戏称为“正太”，但它其实可以称作统计学里的“镇学泰斗”。抛开它上述的众多用途，由于正态分布特有的一些数学性质，使得它在很多统计理论中都占有十分重要的地位。正态分布是很多概率分布的极限分布，其他一些分布的概率（如二项分布）可由正态分布来近似计算，统计推断中许多重要的分布（如 χ^2 分布、T 分布、F 分布）都是在正态分布的基础上推导出来的。

- χ^2 分布：若 n 个相互独立的随机变量均服从标准正态分布，则这 n 个服从标准正态分布的随机变量的平方和构成一组新的随机变量，其分布规律称为 χ^2 分布。

- T 分布：T 分布和正态分布的联系更为直接。上文曾对 U 变换做了解释，当时 U 统计量的构造使用的都是总体的均值、总体的方差。但是，总体参数往往是未知的，如果把总体参数换成样本参数，此时的 U 统计量就变成了 $\frac{x-\bar{x}}{s}$，这就构成了 T 变化，也被称作 T 统计量，此时这个变量的分布就成为了 T 分布。

- F 分布：X、Y 为两个独立的随机变量，如果 X 服从自由度为 k_1 的 χ^2 分布，Y 服从自由度为 k_2 的 χ^2 分布，那么这两个独立的 χ^2 分布被各自的自由度除以后的比率就是 F 分布。

同时，二项分布、泊松分布和正态分布三者之间有着千丝万缕的联系。

- 可以用正态分布近似计算二项分布。

二项分布有两个参数：n 表示试验次数；p 表示一次试验的成功概率。如果一组数据服从二项分布，则可以记为 $X \sim B(n,p)$。在实际运用中，当 n 较大时，一般都用正态分布来近似计算二项分布，此时，$\mu = np$，$\sigma^2 = np(1-p)$。在用正态分布来计算二项分布时，要求 p 不能太接近 1 或 0，除非 n 特别大，否则二项分布会呈现偏态，此时若用正态分布来计算就会有较大误差。但是，如果同时 np 又比较小（比起 n 来说很小），即 np 或者 $n(1-p)$小于 5，那么用泊松分布近似计算更简单一些，毕竟泊松分布跟二项分布一样都是离散型分布。

- 也可以用二项分布来趋近正态分布。

如果 np 存在有限极限 λ，则这列二项分布就趋于参数为 λ 的泊松分布；反之，如果 np 趋于无限大（假设 p 是一个定值），则根据棣莫弗-拉普拉斯定理（De'Moivre-Laplace）中心极限定理，这列二项分布将趋于正态分布。

长得如此可爱（分布图），有那么多实用功能（被广泛应用），还和如此多分布有关联的“正太”，是不是让你也对它刮目相看了呢？对于刚才那些分布之间的相互联系，你是否有一种似懂非懂的感觉？为什么在介绍正态分布、二项分布、泊松分布之间的关联时都有“n 越大越……”这样的句型呢？这个样本量 n 的大小到底与正态分布有什么联系呢？想要知道这些，就得好好看看两位“大叔”是如何牵线搭桥的。

☆本章知识点补充

一个总体参数的区间估计表

参数	点估计量	标准误差	（1–α）的置信区间	假定条件
μ 总体均值	$\bar{x}$	$\frac{\sigma}{\sqrt{n}}$	$\bar{x}\pm z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$	（1）σ已知 （2）大样本（$n\geqslant 30$）
			$\bar{x}\pm z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$	（1）σ未知 （2）大样本（$n\geqslant 30$）
	$\bar{x}$	$\frac{\sigma}{\sqrt{n}}$	$\bar{x}\pm t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$	（1）正态总体 （2）σ未知 （3）小样本（$n<30$）
π 总体比率	p	$\sqrt{\frac{\pi(1-\pi)}{n}}$	$p\pm z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$	（1）二项总体 （2）大样本（$n\geqslant 30$）
σ^2 总体方差	s^2	（不要求）	$\left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}},\frac{(n-1)s^2}{\chi^2_{\frac{1-\alpha}{2}}}\right)$	正态总体

同时附上常用置信水平的 $z_{\frac{\alpha}{2}}$ 值，即可得出下表，这样在求区间估计时就方便多了。

置信水平	α	$\frac{\alpha}{2}$	$z_{\frac{\alpha}{2}}$
90%	0.1	0.05	1.645
95%	0.05	0.025	1.96
99%	0.01	0.005	2.58

那么，普通的正态分布函数值如何求得呢？

若 $Z \sim N(0,1)$，用 $\Phi(z)$ 表示 Z 的分布函数，则它具有如下性质或结论：

- $\Phi(-z)=1-\Phi(z)$
- $P(a \leqslant z \leqslant b)=\Phi(b)-\Phi(a)$
- $P(|z| \leqslant a)=2\Phi(a)-1$

利用这些公式，再加上 U 变换，就能方便地通过标准正态分布表来获取正态分布函数值。

第 5 章

当小“正太”遇上“大叔”——大数定律和中心极限篇

既然“正太”处处都是宝，那让芸芸众生都变成“正太”，整个世界不就和谐美好了吗？如果真能那么容易地规定世间万物都服从正态分布，那么你今天也不会在概率论和统计学的教科书中看到其他分布了，你也不用再为那一堆繁复的分布推导而头昏脑胀了。为了让正态分布的优点普惠众生，统计学家可谓煞费苦心，有几位“大叔”做出了突出贡献，他们推导出了统计学中最负盛名的两大定律——大数定律和中心极限定理。本章我们就来看看站在小“正太”身后强大的智囊团里都有哪些“怪叔叔”。

5.1 正态分布的“左膀”——大数定律

大数定律是什么定律？简单来说就是，随着样本数的增大，用样本的平均数来估计总体的平均数。

通常来说，定律是为实践和事实所证明的，反映事物在一定条件下发展变化的客观规律的论断，如牛顿运动定律、能量守恒定律、欧姆定律等。但是大数定律并不是经验规律，而是严格证明了的定理。

1．伯努利大数定律

首先介绍最为常用、最为有名的大数定律——伯努利大数定律。看名字就知道，推导出该定律的那位“大叔”就是雅各布·伯努利。

瑞士的伯努利家族声名赫赫，他们在数学、科学、技术、工程，乃至法律、管理、文学、艺术等方面都享有盛誉。

我们还是回过头来说说伯努利大数定律是怎么得出的。这就要从他那本神奇的书——《猜度术》说起。《猜度术》开篇第一卷就是《论赌博的计算》，而且在这本书中，伯努利引出了概率论及统计学中的第一个极限定理——伯努利定理，这也是大数定律的雏形。

大数定律可以大致理解为这样一条定律：随机变量序列的算术平均值向随机变量各数学期望的算数平均值收敛。那伯努利在《猜度术》一书中到底提出了怎样的大数定律？伯努利的大数定律是这样描述的：我们来做伯努利试验（伯努利试验是指试验只可能有两种结果：“A”和“非 A”），当试验次数足够大时，某个事件发生的频率将几乎接近于其发生的概率（频率→概率，频率是试验所得，概率是理论期望所趋），表明了频率的稳定性。其表达式如下：

$$\lim_{n\to\infty} p\left(\left|\frac{\mu n}{n}-p\right|<\varepsilon\right)=1$$

将该定律放到现实生活中，最简单的一个例子就是抛硬币。

2．泊松大数定律

与伯努利不同，泊松研究概率论是为了解决法庭审判问题。1837年，泊松出版了专著《关于刑事案件和民事案件审判概率的研究》。泊松在概率统计领域的卓越成就有很多，比如以他名字命名的分布——泊松分布，他也推广了伯努利大数定律。

泊松的大数定律是这样描述的：假如有一组随机变量 $x_1,x_2,\cdots,x_n$，它们之间两两相互独立，同时满足 $p(x_n=1)=p^n$，$p(x_n=0)=q^n$，$p^n+q^n=1$，那么就可以说 $x_1,x_2,\cdots,x_n$ 服从泊松大数定律。其数学表达式如下：

$$\lim_{n\to\infty} p\{\left|\overline{x_n}-\overline{p_n}\right|\geqslant\varepsilon\}=0$$

泊松在他的论文《关于判断的概率之研究》中给出了此大数定律。一开始，该定律只是作为伯努利二项式定律的近似而导出的，但现在它已成为分析放射现象、运输量及一般分布等问题的基础。伯努利大数定律揭示了频率的稳定性，而泊松大数定律则告诉我们，不管独立的随机试验条件如何变化，频率的稳定性这一特征始终不变。

3．切比雪夫大数定律

切比雪夫大数定律是这样描述的：假设有一组随机变量 $x_1,x_2,\cdots,x_n$，它们两两之间相互独立，但必须满足 $E(x_i)=\mu_i$，且 $\mathrm{Var}(x_i)=\sigma_i^2<\mathrm{C}$（这里的 C 是常数）。随机取一个正数 ε，则有 $\lim_{n\to+\infty} p\{\left|\frac{1}{n}\sum_{i=1}^{n}x_i-\frac{1}{n}\sum_{i=1}^{n}\mu_i\right|<\varepsilon\}=1$。如果 $x_1,x_2,\cdots,x_n$ 这组变量具有相同的期望，则可以将这个公式化简为 $\lim_{n\to+\infty} p\{\left|\frac{1}{n}\sum_{i=1}^{n}x_i-\mu\right|<\varepsilon\}=1$。

在切比雪夫大数定律中可以看到，其对随机变量的期望和方差都有要求，但它并不要求随机变量同分布，也就是说，如果有一组随机变量，它的分布情况未知，只要利用它的期望和方差就可以对它的概率分布进行估计，这样反而使伯努利大数定律成为它的一种特殊形式。如果有 n 个相互独立的具有相同期望和方差的随机变量，当 n 趋向于无穷大时，它们的算术平均数几乎就变成了一个常数，这个常数就是它们的数学期望。

4．辛钦大数定律

辛钦大数定律是这样描述的：以切比雪夫的那组随机变量为例，如果这组随机变量符合辛钦大数定律，只需有数学期望并且独立同分布。其数学表达式如下：

$$\lim_{n\to\infty} p\left\{\left|\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\right| < \varepsilon\right\} = 1$$

5.2 正态分布的“右臂”——中心极限定理

中心极限定理是统计领域极为重要的定理，而且和大数定律一样，其背后也有很多学者在为其发展贡献力量。第一条中心极限定理是由棣莫弗-拉普拉斯共同提出的。

1．棣莫弗-拉普拉斯中心极限定理

中心极限定理最早由是法国数学家棣莫弗于 1733 年发现的，他在论文中使用正态分布估计了大量抛掷硬币出现正面次数的分布。这在当时并没有引起人们的关注，直到法国数学家拉普拉斯在 1812 年发表的著作中扩展了棣莫弗的理论，指出二项分布可用正态分布逼近。但同棣莫弗一样，拉普拉斯的发现在当时的学术界也没有引起轰

动。直到 19 世纪末，中心极限定理的重要性才被世人所知。1901 年，俄罗斯数学家李雅普诺夫用更普通的随机变量定义中心极限定理，并在数学上进行了精确的证明。

棣莫弗-拉普拉斯联合推导出的定理可以这样描述：当 n 趋向无穷大时，参数为(n,p)的二项分布将以均值为 np、方差为 $np(1-p)$的正态分布为极限，也就是说二项分布的极限分布是正态分布。其数学表达式如下：

$$\lim_{n\to\infty} p\{\frac{y_n - np}{\sqrt{np(1-P)}} \leqslant x\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} \mathrm{e}^{\frac{-t^2}{2}}\,\mathrm{d}t = \Phi(x)$$

可以看到，该等式的右边是正态分布的分布函数，**所以说中心极限定理研究的是分布的收敛性。**

2．列维-林德伯格中心极限定理

辛钦在大数定律的基础上进一步研究了中心极限定理，他的极限定理经常被用到抽样调查中。但是作为独立同分布要求下的中心极限定理，列维-林德伯格中心极限定理更为有名。

1920 年，林德伯格发表了他的中心极限定理的第一篇论文。这篇论文与李雅普诺夫所做的工作相似，但二人的研究方法不同：林德伯格是基于卷积定理，而李雅普诺夫用的是特征函数。两年后，林德伯格用自己的方法又获得了更稳定的结果，即所谓的 Lindeberg 条件。

中心极限定理是这样描述的：只要一组随机变量独立同分布，不管分布是什么，只要有数学期望μ和方差σ^2，且样本容量 n 足够大，那么样本平均数就趋于期望值为μ和方差为 $\frac{\sigma^2}{n}$ 的正态分布。

这个定理看似简单，但其实作用非常大，我们常说为什么抽样的结果具有可信度，就是因为有这个强大的定理作支撑。在参数统计中，众多理论都是建立在正态分布上的，如参数估计、参数检验等。那么，n 需要多大呢？一般来说，理论上认为 n=30 时就近似正态分布了，n 越大越接近正态分布。当 n=30 时，数据分布是怎样的呢？请看图 5.1。

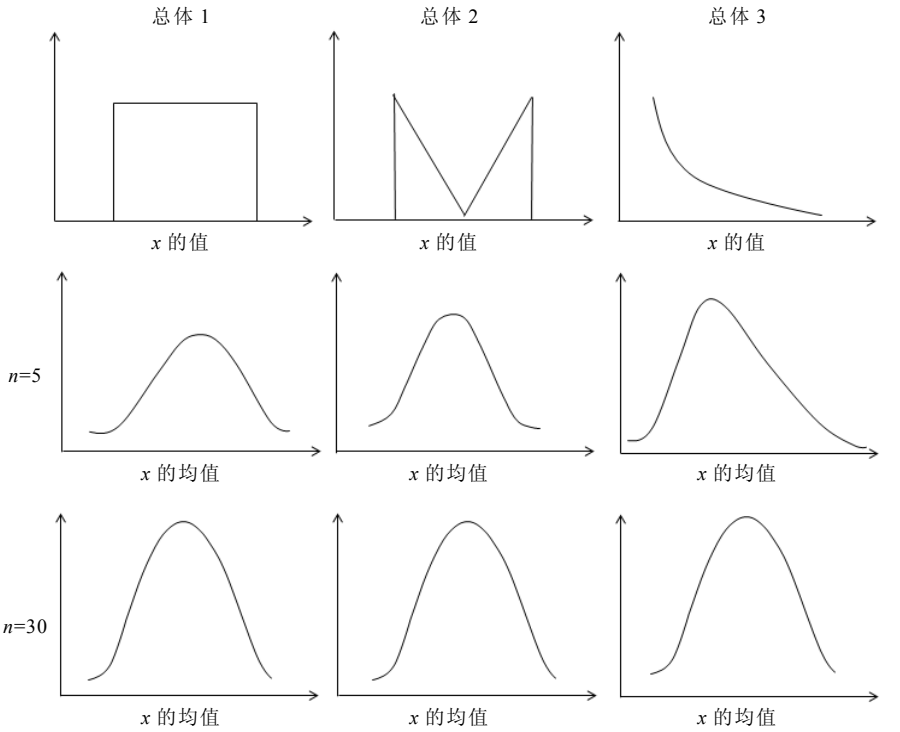


图 5.1　n 取不同值时的数据分布

3．李雅普诺夫中心极限定理

李雅普诺夫从特征函数出发，用一个全新的角度去考察中心极限定理，他发现，中心极限定理在更宽泛的条件下也可以成立。

李雅普诺夫的中心极限定理不要求同分布，只要随机变量是相互独立的，存在期望和方差，那么，当 n 趋向无穷大时，这些随机变量之和的标准化变量的极限就服从标准正态分布。

这里来简单说说数据的标准化。对于数据来说，往往会存在量纲不同的问题，此时最常用的方法就是对其进行标准化处理。但标准化处理也有很多种方法，比如用 log 函数转换、使用 min-max 标准化等，不过这里的标准化处理使用的转换函数，类似于 U 变换。

李雅普诺夫的中心极限定理认为，不管随机变量是什么分布，如果有一个量是由大量相互独立的随机因素影响所造成的，而每一个影响因素在总影响中所起的作用都不是很大，那么这个量就服从或近似服从正态分布。

大数定律和中心极限定理的区别与联系如下。

- 大数定律主要负责阐明大量重复试验的平均结果具有稳定性，解决了变量均值的收敛性问题。
- 中心极限定理主要负责说明随机变量（试验结果）之和逐渐服从某一分布，解决了分布的收敛性问题。

二者之间通过大样本（n 趋向于无穷大）进行联系。更进一步，中心极限定理不仅给出概率的近似表达式，而且能保证它的极限是 1。

5.3 如何牵手“大叔”和“正太”

1. 高尔顿钉板

高尔顿钉板是高尔顿为了研究随机过程给自己做的玩具，类似图 5.2。

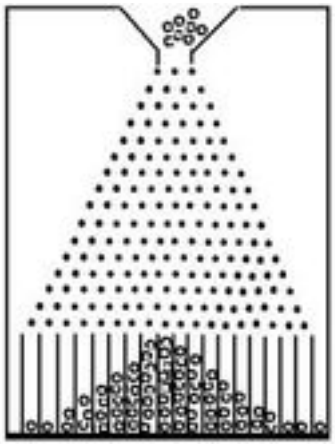


图 5.2 高尔顿钉板

图 5.2 中的每一个黑点表示钉在板上的一颗钉子，它们彼此之间的距离相等，上一层的每一颗钉子的水平位置恰好位于下一层的两颗钉子中间。

这个板和我们要说的主题有什么关系？看到下面那一堆圆球了吗？我们在入口处放进一只直径略小于两颗钉子之间的距离的小圆球，小圆球在下落过程中会碰到钉子，然后以 1/2 的概率向左或向右滚下，接着又碰到下一层钉子……如此继续下去，直到滚到底板的一个格子内为止。一只小圆球落到底板的哪个格子内是随机的，但是当我们把许许多多同样大小的小圆球不断从入口处放下的，只要小圆球的数目足够大，你就会惊喜地发现，它们在底板将堆成近似正态分

布密度函数的图形。这其实就是一个用实物具体模拟中心极限定理的过程。

2．你还在为中奖发愁吗

说回正题，既然大数定律和中心极限定理如此神奇，那么，能否让它们来帮助我们预测彩票号码呢？

其实买彩票就像玩一个数字游戏，摇出号码的过程是随机的，就像抛硬币的过程也是随机的，不过次数多了就会得出一个近似概率；相应地，奖号摇多了，自然也会形成相应的概率分布。如果你是彩民，那么一定听说过冷门号码和热门号码，中奖与否就掌握在“一冷一热”之中，而如何选择“冷热”，其艺术就在大数定律和中心极限定理之中。

随着开奖号码增多，各个号码的出现次数会不断趋近平均值，如果你有足够的时间和精力搜集数据，你就能慢慢地计算出各号码出现的频次和各个号段出现的频次，通过统计计算或许还能找出各个号码之间的关联，从而大大提高中奖概率。

☆本章重点归纳

1．大数定律分类

- 强大数定律：柯尔莫哥洛夫大数定律；博雷尔大数定律。
- 弱大数定律：切比雪夫大数定律；辛钦大数定律；伯努利大数定律。

2．如何区别——以收敛方式来判断

弱大数定律是“以概率收敛”，而强大数定律是“几乎确定收敛

或以概率为 1 收敛、几乎处处收敛”。满足强大数定律的必定满足弱大数定律，反之不成立。

那么，问题来了：什么是以概率收敛？什么是处处收敛？

以概率收敛其实很复杂，举个例子来说明一下。比如我们用拼音打字，一开始总是一个字母一个字母地按，速度很慢。但当我们熟悉了键盘字母的分布后，就可以开始所谓的“盲打”了。不过此时我们还是会打错字母，但随着熟练度不断提高，正确率终将趋于 100%，不过还是会出现几次失误。这就是“以概率收敛”的概念。那么几乎处处收敛呢？那就是没有失误的机会。

3．中心极限定理的分类

一组随机变量相互独立
- 只要独立→李雅普诺夫中心极限定理
- 独立同分布→列维-林德伯格中心极限定理
- 独立同（二项）分布→棣莫弗-拉普拉斯中心极限定理

第 6 章

相关和因果切莫傻傻分不清楚

数据之间的关系应该是推断统计首先需要判断的。佛曰，善恶皆有报，世间万物皆逃不出因果轮回。可是，孰因孰果在这轮回间愈加难辨。但是，自古以来，无论是东方文明古国还是西方各帝国，都没有放弃过对事物因果的研究。在哲学领域里，因果论也得到众多哲学家的追捧。当然，最后谁也没有把真正的世间因果说清楚。既然不能严格证明事物间的因果，那么用“相关”来描述事物间的关系不就可以了吗？高尔顿在研究回归分析的时候就提出了“相关”这一概念，主要用来区分回归平均值的直线是否是一条清晰的直线。

不过，在没有说明什么是回归分析之前，先来说说统计学家为何会引入“相关”这一概念。

6.1 为了“不确定”的确定

在统计学里，与相关关系真正对应的是函数关系而非因果关系。二者最大的区别在于它们刻画“关系”时是否能够用一个明确的数学式子来表示。

先来说说函数关系。函数关系是两个变量（或者多个变量）之间有完全确定的关系。我们一般会将变量设置为自变量和因变量，当自变量给定时，若存在函数关系，则由各个自变量构成的函数具有确定的函数值。而如果仅仅有相关关系，多个变量之间并没有严格的确定关系，那么当一个变量（或者多个变量）变化时，另一个变量的取值就会有一定的随机性。如果我们仔细留意身边的一些事物，就会发现，我们的身边到处充斥着“相关”的身影。

比如，我们经常会在浏览网页时遇到一些推送信息。我们在豆瓣上看了一篇影评，下方就会有与之类似的影评；同样，当我们通过点评类网站选择餐厅的时候，也经常会有相关餐厅推荐……这些都是一种相关关系——事务之间都有共同点，但不可能完全一样，这其中含有一定的随机因素。

再回到统计学上，我们在研究一些变量数据的时候，可能无法找到一个函数能够完美地刻画数据分布情况。于是，统计学家设计出了很多种统计方法和统计量来描述这种不确定的关系。

6.1.1 散点图

很多人在拿到数据后会先简单地画一幅散点图，了解一下数据的大致情况，如图 6.1 所示。

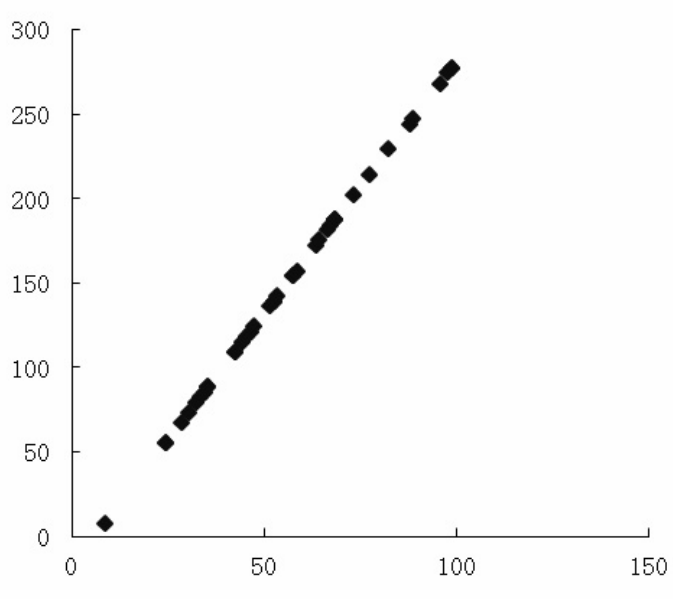
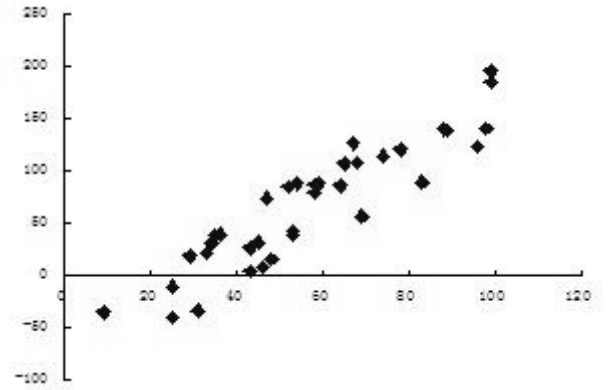


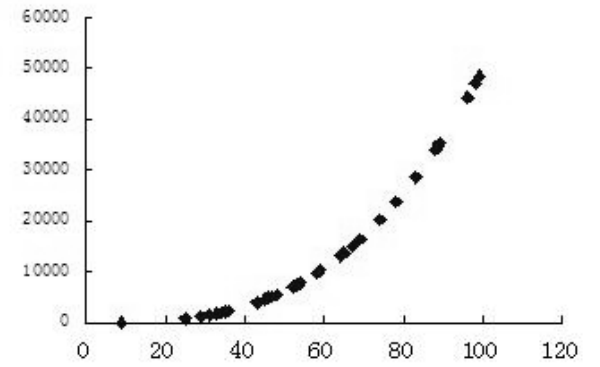
图 6.1 理论假设下的散点图

如果你在作图的时候得到的是类似图 6.1 的散点图，那么恭喜你，这批数据应该是相当让人满意的，因为图中的数据分布几乎为一条直线，这样的数据分布意味着数据之间存在线性关系，更确切地说是存在函数关系，可以很容易地通过统计建模的方法得到数据模型。

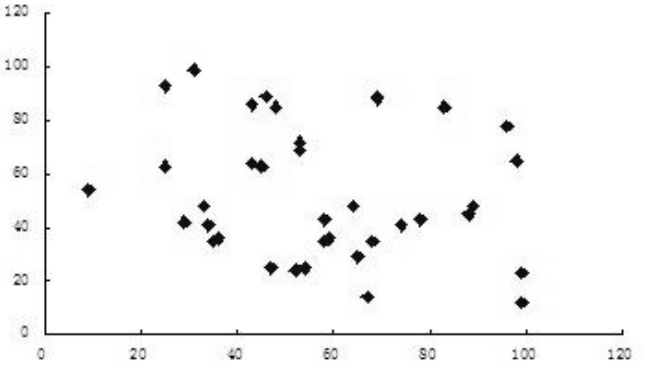
不过，上述情况只是理论假设，大多数时候的散点图类似图 6.2。



（a）



（b）



（c）

图 6.2　一般情况下的散点图

首先来看图 6.2（a）。通常情况下，在做一些回归分析的时候此类散点图最为常见，因为这类散点图描绘了数据之间的线性相关性，很适合用来建立线性模型。从图上看，两个变量基本是正向变动的。也就是说，当 X 轴的数值增加时，Y 轴的数值也相应增加，只不过每个点增加的幅度不尽相同。与该图对应的是负向变动，此时 X 轴的数值增加（减少），Y 轴的数值相应减少（增加）。不管怎样，它们都属于便于分析建模的数据集种类。

再来看图 6.2（b）。可以看出，它不能算是一条直线，所以一般不会用来描述线性相关性，它所刻画的是数据之间的曲线相关性，两个数据的关系非常密切。对于明显的曲线关系，可以通过线性组合变化使得数据符合线性相关。

如果散点图如图 6.2（c）所示，那么这样的散点图通常预示着两个变量基本不相关。

有人说，读图识数来做判断太过主观，为了证明统计学是一门经得起推敲的学科，统计学家经过层层验算，终于找到了描述数据之间相关性的方法，更准确地说是描述数据之间线性相关性的方法。

6.1.2 相关系数

相关系数的诞生正是为了弥补（或者说解决）散点图主观、模糊的不足之处。下面介绍三个相关系数的优缺点。

1．皮尔逊相关系数

皮尔逊相关系数是最为常用的相关系数，又称皮尔逊积矩相关系

数，在诸多软件中，它时常被写为“Pearson 相关系数”，这是一个用于度量基于正态分布的定距变量间的线性相关关系的数值。那么，问题来了，什么是定距变量？

定距型数据是数字型变量，可以求加、减、平均值等，但不存在基准 0 值。也就是说，当变量值为 0 时不是表示没有，比如当温度为 0℃时，不能说没有温度，温度就是一个定距变量。

再回到相关系数。一般而言，简单相关系数可以用 ρ 来表示（有时也会用 r 来表示），其计算公式为：

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

上述公式很简单，可以看作将两组数据首先做 Z 分数处理，然后将两组数据的乘积和除以样本数。前面说过，皮尔逊相关系数是一个基于正态分布的线性相关系数，所以它具有如下几个约束条件：

- 两个变量之间有线性关系。
- 变量是连续变量。
- 变量均符合正态分布，且二元分布也符合正态分布。
- 两个变量之间相互独立。

仅给出相关系数的计算公式还不够，问题是得出这个数值有什么用？要回答这个问题，就要说到相关系数的取值范围。无论是样本的

相关系数还是总体的样本系数，它的取值范围都为[–1,1]。具体说来，可以根据相关系数的数值作如下判断。

- 相关系数有正有负，相关系数为正时表明数据呈线性正相关；相关系数为负时则表明数据呈线性负相关；相关系数为零则表明数据之间并无线性相关关系。
- 不同的相关系数取值代表不同的相关程度。
  - 若$|\rho|$>0.95，则变量间显著线性相关。
  - 若 0.8≤$|\rho|$≤0.95，则变量间高度线性相关。
  - 若 0.5≤$|\rho|$<0.8，则变量间仅中度线性相关。
  - 若 0.3≤$|\rho|$<0.5，则变量间低度线性相关。
  - 若$|\rho|$<0.3，则认为变量间不相关。

有了这些判断系数参考值，要判别两个变量的线性关系就方便多了。不过，在皮尔逊相关系数中有一个约束条件是：变量是服从正态分布的连续变量。虽说很多数据在大数定律和中心极限定理下近似服从正态分布，但还是有特殊情况，怎么办？

2．斯皮尔曼等级相关系数

斯皮尔曼等级相关是根据等级数据来对两个变量之间的相关关系进行研究的方法。它是依据两列成对等级的各对等级数之差来进行计算的，所以又称为“等级差数法”。

斯皮尔曼等级相关系数对数据条件的要求没有皮尔逊相关系数那么严格，只要两个变量的观测值是成对的等级评定资料，或者是由连续变量观测资料转化得到的等级评定资料，不论两个变量的总体分布形态、样本容量的大小如何，都可以用斯皮尔曼等级相关系数来进行研究。

斯皮尔曼等级相关系数的计算公式要比皮尔逊相关系数简单多了，只有两个参数：n，表示等级个数；d，表示两列成对变量的等级差数。具体的计算公式如下：

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

不过，计算该相关系数需要三个步骤；

（1）把数量标志和品质标志的具体表现按等级次序编号（也就是我们常说的排序）。

（2）按顺序求出两个标志的每对等级编号的差。

（3）按公式计算相关系数。

斯皮尔曼同样适用于皮尔逊相关系数的判断法则，取值范围都是[−1,1]，不过二者也是有区别的。除了斯皮尔曼相关系数对分布、变量连续与否没有要求外，斯皮尔曼相关系数由于使用的等级数据（更专业的叫法为秩统计量）属于非参数相关系数，假如两个变量之间具有单调的函数关系，那么就可以认为这两个变量是完全相关的（$|\rho|=1$），这与皮尔逊相关性不同，后者只有在变量之间具有线性关

系时才是完全相关的。

3．肯德尔等级相关系数

肯德尔系数是由林德伯格发明的。肯德尔系数有好几个，常用的有适合两列等级变量相关度衡量的 τ 系数，由于它和斯皮尔曼相关系数对数据的要求一致，所以不作介绍，来说说不一样的，如肯德尔和谐系数 W。

肯德尔和谐系数是表示多列等级变量相关程度的一种方法，更通俗地说，它衡量的是变量间是否一致性的问题。

具体来说，就是让 k 个评价者对 n 件事物或 n 种作品进行等级评定，每个评价者都能对 n 件事物（或作品）的优劣、喜好、大小、高低等排出一个等级顺序。因此，最小的等级序数为 1，最大的为 n，这样 k 个评价者便可得到 k 列从 1～n 的等级变量数据，这是一种情况。另一种情况是一个评价者先后 k 次评价 n 件事物或 n 件作品，也采用等级评定的方法，这样也可得到 k 列从 1～n 的等级变量数据。然后对这类 k 列等级变量综合起来求相关，此时就用到了肯德尔和谐系数 W。肯德尔和谐系数的计算公式如下：

$$W=\frac{s}{\frac{1}{12}k^2(n^3-n)}$$

其中，s 定义为：

$$s=\left(\sum R_i-\frac{\sum R_i}{n}\right)^2$$

式中涉及三个参数：R_i为每件被评价事物的 k 个等级之和；n 为被评价事物的件数，即等级数；k 为评价者的数目或等级变量的列数。

肯德尔和谐系数基于这样一种思想：如果各列变量完全一致，那么各被评价的事物（或人），其各评价者所评的等级应该相同；如果评价的等级不同，则 s 变小，一致性程度降低，如果完全不相关，则所评各等级之和应该相等，其最大可能方差 s 应该为 0，这样等级和的方差与最大可能的方差的比值便是和谐系数。从公式中可以推导出其值必定介于 0 与 1 之间。

上述几种方法测算出的两个变量间的相关度都称为单相关系数，除此之外，还有复相关系数。复相关系数是测量一个变量与其他多个变量之间线性相关程度的指标，也就是某一变量与其他变量线性组合后预测值的相关度。比如，想要了解立定跳远的距离和年龄、身高、体重、性别等多个因素之间的关系，就需要使用复相关系数。对于复相关系数，我们需要记住一点，虽然它也称为相关系数，但是它的取值范围为[0,1]。

虽然散点图和相关系数的诞生为确定数据之间的不确定关系提供了一把解开迷雾之锁的钥匙，但无论是科学领域还是哲学领域，人们都没有放弃研究数据之间的确定关系。在长达几个世纪的角逐中，哲学家、科学家都对因果论是否存在、如何存在产生了激烈的辩驳，下面来看看统计领域如何看待因果关系。

6.2 上帝掷骰子

在量子力学里，爱因斯坦有一句名言：“上帝永远不掷骰子。”这句话来源于他与波尔的一场争论，同时也是针对海森堡测不准理论而言的，大致是说世间万物是不存在随机性的，一切都是确定的，我们可以用科学来严格推算出一切。不过，现今的量子力学基本证实了“上帝是掷骰子的”，霍金更是说：“上帝不光掷骰子，还把骰子掷到我们看不见的地方。”这句话说的就是世间不仅存在随机性，而且这些随机性可能无法被观测到。

通过这个例子告诉大家，科学理论并非简单的“因为……所以，科学道理”，很多时候我们看到的貌似“因果”也许并非因果关系。

1．相关≠因果

给出如下几句话，大家来进行判断：

（1）越是成功的人，睡眠越少。

（2）努力学习，成绩就会好。

（3）读了 MBA，就能当 CEO。

（4）吸烟导致肺癌。

这 4 句话会在生活中的各个角落里不经意地出现。这几句话貌似正确，实则不然。睡眠时间和成功与否其实并没有因果关系。虽然我们将成功与否与睡眠时间进行量化分析确实可以得到一定的相关性，但睡眠时间并不是成功与否的决定性因素。

所谓因果关系，是指某个因素的存在一定会导致某个特定结果的产生。简单地说，就是 A→B，即事件 A 的发生必然会导致事件 B 的发生。而上述几个命题并不符合因果关系的定义。

为什么我们总是不经意间将原本仅是相关关系的事物发展为因果关系呢？除去我们固有的追根溯源的心理状态外，其实很多时候我们得出错误的因果关系还可能源于以下几种情况：

（1）胡乱确定因果关系。有个古老的谬误：“如果 B 紧跟着 A 发生，那么 A 一定导致 B。”在这里，或许 A 是 B 的因，B 是 A 之果，但更可能的情况是，A 和 B 并不互为因果，而都是第三种因素的产物。

（2）小样本错误。这是一种数据“陷阱”，原因在于采样过少，即使分析和推理过程正确，也不一定能得出正确的结论。

（3）把相关关系当作因果关系。在很多情况下，变量之间只存在相关关系，是否存在因果关系仍然是个未知数。因此，在明确变量之间确实存在因果关系之前，不宜盲目下结论。

2．此因果非彼因果

那是不是我们真的对因果关系的验证没有办法了呢？也不是。在统计和计量经济领域有一种因果检验方法叫作格兰杰因果检验（Granger Causality），是经济学家克莱夫·格兰杰所创。这是否意味着我们可以用科学的方法来论证因果关系了呢？这并不是那么容易的事，其实此因果非彼因果。

克莱夫将自己的检验定义为：依赖于使用过去某些时点上所有信息的最佳最小二乘法预测的方差。

这个检验其实并非广义上的因果检验，其应用有很大的局限性。首先，它仅适用于平稳的时间序列；其次，它的运用依赖于变量的过去信息（也叫滞后信息）。一种对格兰杰因果检验更为具体的定义是：如果在包含了变量 x、y 过去信息的条件下，变量 x 有助于解释变量 y 的将来变化，则认为变量 x 是引起变量 y 的格兰杰原因。

这个定义太过轻巧，所以，在通常情况下，格兰杰检验所得的因果关系都定义为统计意义上的因果关系，若去掉“统计”两个字，那它其实就和因果没有多少关系了。在格兰杰描述中有一个很重要的词汇——预测。正是这个词道出了它不同于我们通常理解上的因果关系之间的区别所在。格兰杰的因果检验考察的是变量的预测性；而我们通常所说的因果其实是逻辑意义上的。

说了这么多关于“关系”的内容，就是希望大家能够谨慎地对变量之间的关系下定论。相关也好，因果也罢，对于数据，我们要抱以敬畏之心，计算检验时也需要选择恰当的方法，因为何种“关系”将会成为我们后续分析的基础假设。

☆本章知识拓展

说了那么久的相关关系，对各类相关系数也做了虽简单但基本全面的介绍，在本章的结尾，我们来对两个主要的相关系数——皮尔逊相关系数和斯皮尔曼相关系数的优缺点做一个总结。

1．皮尔逊相关系数

- 优点：
  - 数学表达式简单。相关系数是一个介于–1～1 的常数，有了数值便可以直接量化数据的相关度。
  - 相关系数不受变量单位的限制。即可以计算身高（cm）和体重（kg）之间的相关系数，也可以计算小学生学习成绩（分数）和看电视时长（小时）之间的相关系数。
- 缺点：它接近于 1 的程度与数据组数 n 相关，这容易造成一种假象。因为当 n 较小时，相关系数的波动较大，对有些样本来说，相关系数的绝对值易接近于 1；当 n 较大时，相关系数的绝对值容易偏小；特别是当 n=2 时，相关系数的绝对值总为 1。因此，在样本容量 n 较小时，仅凭相关系数较大就判定变量 x 与 y 之间有密切的线性关系是不恰当的。

2．斯皮尔曼相关系数

- 优点：适用范围广泛，斯皮尔曼相关系数对数据条件的要求没有皮尔逊相关系数那么严格，只要两个变量的观测值是成对的等级评定资料，或者是由连续变量观测资料转化得到的等级评定资料，不论两个变量的总体分布形态、样本容量的大小如何，都可以使用斯皮尔曼相关系数来进行研究。
- 缺点：一组能用积差相关计算的数据，如果改用等级相关，精确度会低于积差相关。因此，凡符合积差相关条件的，最好不要用等级相关计算。

最后附上使用 SPSS 软件计算相关系数的方法：

打开 SPSS 软件后，首先导入数据，然后在工具栏处依次单击“分析”→“相关”→“双变量”（“双变量”里有皮尔逊相关系数、肯德尔和谐系数、斯皮尔曼相关系数），将两个变量放入“变量”框中。

接下来，在“相关系数”框中选择“Pearson”。为什么选择“Pearson”呢？其实这是由数据类型来决定的。如果变量为连续性变量，则可采用皮尔逊相关分析；如果为分类变量，或者一个分类变量、一个连续性变量，则可以用斯皮尔曼相关分析。

选择好变量后，如果需要对数据进行一定的描述，则可以单击右上角的“选项”按钮。如果需要进行描述性分析，则可以选择均值和标准差，分别对数据的大小和离散程度做出一定的描述，并单击“确定”按钮。

之后，等待结果即可。

第 7 章

“小”亦可为，“大”而佐之

上一章我们通过相关分析，初步踏入了推断统计学的大门，本章主要介绍推断统计学里两个非常重要的估计方法——最小二乘估计法和最大似然估计法。依照从“小”到“大”的顺序，先来说说最小二乘估计法的思想原理和发展来源。

7.1 这个“小二”一点都不“二”

对于“二乘”的概念，其实最早是基于误差研究展开的。关于最小二乘法的记录最早出现在勒让德 1805 年发表的一本著作《计算彗星轨道的新方法》里，但他只提出了一个大概的思路：让误差平方和达到最小，在各方程的误差之间建立一个平衡，从而防止某一极端误差取得支配地位，这有助于揭示系统更接近真实的状态。

勒让德使用的最小二乘法理论思想固然好，但在实际计算时略有不妥之处，经过一番修正后，再来看看现今的最小二乘法到底是如何

计算和使用的。

我们所说的最小二乘法是通过将误差平方和最小化，以此来寻找与现有数据匹配最佳的函数。什么是误差平方和最小化呢？

所谓误差，是指通过得到的数据构建了模型并且拟合出了新的数据，不过这些拟合数据是“计算”出来的，它们与实际数据之间存在一定的偏差，这些偏差就可以理解为误差。而我们要做的就是力求将这些拟合值与实际值之间误差的平方和最小化，这样就能找到一个理想的模型。

下面通过一个简单的例子来一窥“最小二乘法”的真颜。前面我们说过身高和体重之间是有相关关系的，不妨用这两个变量来构建一个一元的线性模型。身高作为自变量 x，体重作为因变量 y，则可以用一个表达式 $y=ax+c+\varepsilon_i$ 来刻画它们之间的关系（这里的 ε_i 就是随机误差项）。

模型的框架已经搭好，接下来要解决的问题就是怎么计算系数 a 和常数 c。“小二”是这么说的：不管你用什么方法求 a 和 c，只能构建一条直线无限逼近你的数据，不可能完全连接所有的观测值。虽然穿过这些观测值的直线有无数条，但我们要找的就是直线拟合值与实际观测值误差最小的那条，也可以理解为是误差平方和最小的那条，用数学公式来表达就是：$\min(\sum {\varepsilon_i}^2)=\min\sum(y-\hat{y})^2$。

基于这个思想，可以用最小二乘法推得参数 a 和 c，具体方法如下：

$$\min\sum(y-\hat{y})^2=\min\sum(y-ax-c)^2$$

为了使该值最小，则估计值 $\hat{a}$ 和 $\hat{c}$ 应该满足：

$$\frac{\partial\left(\sum{\varepsilon_i}^2\right)}{\partial\hat{c}}=-2\sum(y-\hat{a}x-\hat{c})=0$$

$$\frac{\partial\left(\sum{\varepsilon_i}^2\right)}{\partial\hat{a}}=-2\sum(y-\hat{a}x-\hat{c})x=0$$

从而得出：

$$\hat{a}=\frac{n\sum xy-\sum x\sum y}{n\sum x^2-(\sum x)^2}$$

$$\hat{c}=\frac{n\sum x^2\sum y-\sum x\sum xy}{n\sum x^2-(\sum x)^2}$$

式中，n 为样本量。

这就是参数 a 和 c 的最小二乘估计值，将这两个参数估计值带入模型就可以得到 $y=\hat{a}x+\hat{c}+\varepsilon_i$ 这条最为“贴合”的直线。不过有的读者可能会问，怎么证明这条直线就是最贴合的那条？

高斯给出了肯定的答案，其判断标准就是著名的“高斯-马尔科夫定理”。该定理的描述如下：在给定的假定条件下，最小二乘估计量是具有最小方差的线性无偏估计量。也就是说，如果我们所要建立的线性模型的经典假设成立，则没必要再去寻找其他的无偏估计量，没有一个比最小二乘估计量更好了。即便有这么一个估计量，它的方差也最多和最小二乘估计量一样小。这样就保证了采用最小二乘法计

算得出的参数，并以之构建的模型是这批数据中最优的模型。

那么问题又来了：什么是无偏估计量？

无偏估计量等于被估计的量的统计估计量。举个例子：$\hat{\lambda}$是λ的一个估计量，如果$E(\hat{\lambda})=\lambda$，那么$\hat{\lambda}$就是一个无偏估计量。

还有一个问题：定理中提到了“在给定的假定条件下”这句话，都有哪些假定条件呢？高斯-马尔科夫定理给出的假定条件有如下几条：

（1）要求所有参数均为常数，这样就保证了模型为线性模型。比如，有一个模型是$y=a+b_1x_1+b_2x_2+\cdots+b_nx_n+\varepsilon_i$，那么根据假定，则$\mathrm{a}, b_1,\cdots,b_n$必须为常数，其中$\varepsilon_i$为误差项。

（2）如果有n个调查样本，那么这n个样本必须是从总体中随机抽取的。

（3）在样本（总体）中，没有解释变量是常数；而且解释变量之间不能存在完全共线性，否则该方程将会无解。

（4）总体方程的误差项均值为 0，并且误差项均值不受解释变量的影响。

（5）误差项的方差不受解释变量影响且为一个固定值（同方差性）。

只要符合这 5 个假定条件，即可放心地使用最小二乘法来估计参数。

7.2 另辟蹊径的最大似然估计法

“最大似然”一词最早出现在高斯的误差正态分布的论证中，不过却得名于费希尔于 1912 年发表的论文。“似然”一词其实是英语“likelihood”的中文翻译，意为“可能性”。所以最大似然估计法是基于概率而言的，也就是说最大概率的估计方法。由此便可以看出，“最小二乘法”和“最大似然估计法”之间的差别在于，前者是基于几何意义上距离最小，而后者是基于概率意义上出现的概率最大。

最大似然原理的基本思想为：当从模型总体随机抽取 n 组样本观测值后，最合理的参数估计量应该使得从模型中抽取该 n 组样本观测值的概率最大，而不是像最小二乘估计法旨在得到使模型能最好地拟合样本数据的参数估计量。

一种想法是：有一个随机试验，这个随机试验可能会有若干个可能的结果。如果只进行一次试验，而结果 A 出现，那么，是不是可以认为试验条件对 A 出现有利，也就是说 A 出现的概率很大？一般地，事件 A 发生的概率与参数 θ 相关，A 发生的概率记为 $P(\mathrm{A},\theta)$，则 θ 的估计应该使上述概率达到最大，这样的 θ 被称为最大似然估计。那么，寻找这个参数 θ 的估计值就成为此方法的重中之重。

引用一个常见的例子：某一天，你的朋友和一位猎人一起外出打猎，此时一只野兔出现，只听“砰”的一声，野兔应声倒下。这一枪是谁放的？有可能是一起放的，那这只兔子是谁打中的？一般来说，猎人打中兔子的概率要高于你的朋友，这其实就是最大似然估计的思想。

那么，最大似然估计又是如何通过概率计算来获得参数估计值的呢？

第一步：构造似然函数。

对于似然函数，首先要查看变量 X 属于离散型数据还是连续型数据。

假设总体 X 是离散型随机变量，其概率函数为 $p(x;\theta)$，其中 θ 是未知参数。现在，假设 $X_1,X_2,\cdots,X_n$ 是来自总体 X 的样本，$X_1,X_2,\cdots,X_n$ 的联合概率函数为 $\prod_{i=1}^{n}p(X_i;\theta)$（$\theta$ 是常量，$X_1,X_2,\cdots,X_n$ 是变量）。

如果已知样本取值是 $x_1,x_2,\cdots,x_n$，那么事件 $\{X_1=x_1,X_2=x_2,\cdots,X_n=x_n\}$ 发生的概率应该为 $\prod_{i=1}^{n}p(x_i;\theta)$，这个概率随 θ 值的变化而变化。按照最大似然估计的思想，既然样本值 $x_1,x_2,\cdots,x_n$ 出现了，它们出现的概率相对来说比较大，那么这就应该使得 $\prod_{i=1}^{n}p(x_i;\theta)$ 取比较大的值。换句话说，θ 应使样本值 $x_1,x_2,\cdots,x_n$ 的出现具有最大的概率。现在可以将上式看作一个包含 θ 的函数，并用 $L(\theta)$ 表示，于是就有 $L(\theta)=L(x_1,x_2,\cdots,x_n;\theta)=\prod_{i=1}^{n}p(x_i;\theta)$，这里的 $L(\theta)$ 就是似然函数。

如果总体 X 是连续型随机变量，它的概率密度函数为 $f(x;\theta)$，假设取得的样本观察值为 $x_1,x_2,\cdots,x_n$，因为随机点 $(X_1,X_2,\cdots,X_n)$ 取值为 $(x_1,x_2,\cdots,x_n)$ 时联合密度函数值为 $\prod_{i=1}^{n}f(x_i;\theta)$，所以按最大似然估计，同样应选择使此概率达到最大的 θ 值，此时的似然函数为

$L(\theta)=\prod_{i=1}^{n} f(x_i;\theta)$。

第二步：求对数似然函数。

既然我们已经找到了似然函数，接下来的事情似乎就有了方向。还记得之前的“最小二乘法”是如何得到估计值的吗？求导！这同样适用于最大似然估计法。只不过在这一过程中还需要多走一步——求似然函数的对数函数。

为什么要求似然函数的对数函数？因为 $\ln L$ 是 L 的增函数，所以 $\ln L$ 与 L 在 θ 的同一值处取得最大值，而取了对数后，一些数学处理就方便多了。那么，在这一步中，只需把似然函数写成对数似然函数 $l(\theta)=\ln L(\theta)$ 即可。

第三步：求导数。

原本最大似然估计法是在参数 θ 的可能取值范围 Θ 内，选取使 $L(\theta)$ 达到最大的参数值 $\hat{\theta}$，以此来作为参数 θ 的估计值。用数学语言来说，那就是选取 θ，使得 $L(\theta)=L\left(x_1,x_2,\cdots,x_n;\hat{\theta}\right)=\max_{\theta\in\Theta} L(x_1,x_2,\cdots,x_n;\theta)$。

因此，求总体参数 θ 的最大似然估计值的问题其实就是求似然函数 $L(\theta)$ 的最大值问题。这可通过解求导方程 $\frac{\mathrm{d}L(\theta)}{\mathrm{d}\theta}=0$ 得到。也可以将求导方程改为 $\frac{\mathrm{d}\ln L(\theta)}{\mathrm{d}\theta}=0$。

此时，只要能够解出方程，即可得到 $\hat{\theta}$，也就是参数 θ 的最大似然估计值。

总结一下求最大似然估计值的步骤：

（1）由总体分布导出样本的联合概率函数（或联合密度）。

（2）把样本联合概率函数（或联合密度）中的自变量看作已知常数，而把参数θ看作自变量，得到似然函数$L(\theta)$。

（3）求似然函数$L(\theta)$的最大值点（常转化为求对数似然函数$L(\theta)$的最大值点）。

（4）在最大值点的表达式中，只要用样本值代入就可得到参数的最大似然估计值。

这里还有一个很重要的问题：什么时候用最大似然估计法，什么时候用最小二乘估计法？

最小二乘估计法是基于拟合最佳出发的，对数据分布没有要求，而最大似然估计法则需要已知概率分布函数。一般地，如果我们的数据满足正态分布函数的特性，则最大似然估计法和最小二乘估计法是等价的。

7.3 他山之石，或可攻玉

当然，要估计参数不可能只有“最大”、“最小”两种方法，只是比它们略有逊色，常用的有以下两种方法。

- **矩估计法**：用样本矩估计总体矩。举个例子，我们常用样本均值估计总体均值，其中就需要大数定律和中心极限定理来作保障。

- **最小一乘法：**要求各实测点到回归直线的纵向距离的绝对值之和最小。

1760 年，博斯科维奇在研究子午线时就提出了最小一乘法。虽然“最小一乘法”和“最小二乘法”都是为了求解方程组的解，但最小一乘法所求出的最佳估计量是一个条件中位数（最小一乘回归直线是中位数直线），而最小二乘法所求出的估计量则是条件均值（最小二乘回归直线是均值直线）。于是引发了均值和中位数如何选择的问题。当然，在数据及误差项服从正态分布的时候二者没有区别，但如果误差不服从正态分布，则最小一乘法的统计性能要优于最小二乘法。

说了那么多估计法，我们好像对此章的“配角”——估计量的关注有些少。其实如何选择合适的估计量也是一个非常值得探讨的话题。可以用来估计未知参数的估计量有很多，而我们想要选择一个优良估计量，则必须对估计量的优良性定出准则，这种准则不是唯一的，在实际操作中，可以根据实际问题和理论研究的方便性进行选择。

优良性准则有两大类：一类是小样本准则，即在样本大小固定时的优良性准则；另一类是大样本准则，即在样本大小趋于无穷时的优良性准则。最重要的小样本优良性准则是无偏性及与此相关的一致最小方差无偏估计，其次有容许性准则、最小化最大准则、最优同变准则等。大样本优良性准则有相合性、最优渐近正态估计和渐近有效估计等。只有选择了良好的估计量，再配以合适的估计方法，才能得到一个符合总体真实情况的参数估计值。

☆本章知识拓展

1809 年，高斯发表了其数学和天体力学的名著《绕日天体运动的理论》。在此书末尾，他写了一节有关“数据结合”的内容，实际涉及的就是误差分布的确定问题。其实说到这个理论，在最大似然函数的知识中已经有所涉及——构造似然函数，高斯就使用这个函数来寻求最适合的误差密度函数。具体做法如下：

先设真值为 θ，n 个独立测量值为 $X_1,X_2,\cdots,X_n$。高斯把后者的概率取为 $L(\theta)=L(\theta;X_1,X_2,\cdots,X_n)=f(X_1-\theta)\cdots f(X_n-\theta)$，其中 f 就是待定的误差密度函数。

接下来，他又提出两个创新的想法。

第一个创新点在于：他不采取贝叶斯式的推理方式，而直接把能使该式达到最大的 $\hat{\theta}$ 作为 θ 的估计，也就是使 $L(\theta)=\max L(\theta)$ 成立的 $\hat{\theta}$。我们知道，$L(\theta)$ 是似然函数，所以把满足该式的 $\hat{\theta}$ 称为 θ 的最大似然估计。

第二个创新点在于：他把问题倒过来，先承认算术平均 $\overline{X}$ 是应取的估计，然后寻找误差密度函数 f 以迎合这一点。最终证明，只有在 $f(x)=\dfrac{1}{\sqrt{2\pi}\sigma}\exp\left(-\dfrac{x^2}{2\sigma^2}\right)$ 条件下才能成立，这里 $\sigma>0$ 为常数。这不就是正态分布吗？

不过，当时很多人对此存在异议，因为高斯的说法有一点循环论证的味道：由于算术平均是优良的，所以推出误差必须服从正态分布；反过来，由后一结论又推出算术平均及最小二乘估计的优良性，故必

须认定这二者之一（算术平均的优良性、误差的正态性）应作为出发点。但算术平均到底有没有自行成立的理由尚无定论，以它作为理论中一个预设的出发点，终觉有其不足之处。不过，后来拉普拉斯最终运用中心极限定理将这个理论断裂的一环连接起来，使之成为一个和谐的整体。

第 8 章

从先放牛奶 or 先放热茶说起

在统计推断中，我们是围绕两个核心问题展开讨论的：一个是参数估计；另一个是假设检验。这两个问题虽然看似独立，但其实它们之间有着千丝万缕的联系。尤其是在回归分析中相遇时，就会很自然地融为一体。但为何说二者是统计的“两重天”，主要还在于它们的侧重点有所不同。

参数估计说的是用样本统计量去估计总体的参数——重在估计；假设检验则先对总体参数提出一个假设值，然后利用样本信息判断这一假设是否成立——重在论证。在假设检验出现在统计学之前，统计学家更多地关注如何估计一些参数，但当费希尔出现后，假设检验却成为统计学一个非常重要的学科理论。

要说假设检验，从字面上就能看出只要是针对你所假设的进行检验就属于假设检验的范畴。一般情况下，学习一门学科总得对它的发展历史有所了解，我们是不是也该说说“假设检验”的历史呢？不过，

这好像真的不太好说。为什么呢？并不是说它没有历史来源，而是这个检验其实很早之前就有人提过，比如阿布兹诺特等人对婴儿性别的检验、众多学者对正态均值的检验等，只不过当时都没有对其正式命名，也没有进行系统的研究。在那个时代，这些人的检验思想和“假设检验”非常接近，但又无法确切地追溯其源。所以，我们不对它的发展史做过多的深究。不过，说起假设检验，有这么几个人还是需要了解的。首先是老皮尔逊和费希尔。“统计之父”老皮尔逊的拟合优度检验和费希尔的显著性检验在当时使得假设检验在统计学中的地位有了明显提升；紧接着，奈曼和小皮尔逊则在制定检验统计量的原则和标准上做了进一步的探讨，建立了一套有效的理论，使得假设检验的理论体系趋于完整。

不过现在提起假设检验，往往指的是费希尔的显著性检验，那为什么老皮尔逊的拟合优度检验相较之下就不受重视呢？其实当时老皮尔逊写了一篇论文，被誉为假设检验的开山之作。可惜，那长长的标题里竟然从未出现过“hypothesis testing”，而更重要的原因在于，正是这篇论文，挑起了费希尔和老皮尔逊的战火，并且费希尔赢了。当然，这不是说老皮尔逊的拟合优度检验不好，事实上，现在的统计学对拟合优度的关注度也很高，而且是目前较为常用的几种检验之一。

回过头来说说费希尔。费希尔最有名的假设检验是均值和回归系数的显著性检验，但除此之外，他也对拟合优度检验做过研究，另一个非常有名的检验就是方差分析。说到这里，先给大家讲个故事：

某天下午，费希尔和一群穿着优雅的女士在花园里边喝英式下午

茶边聊天。这时，一位女士忽然说：只要给我喝一口奶茶，我就能告诉你这杯奶茶是先放奶还是先放茶的。众女士在面露惊叹之余，纷纷要求她详细道来。

此时的费希尔不以为然，嘴角一扬，说道：那我就来做个假设。要分辨出先加奶还是先加茶虽然不是很容易的事，但其实谁都有可能猜对。我先假设这位姑娘是没有这个能力完全分辨出来的。

于是费希尔亲自调配了 8 杯其他条件相同，只是倒茶顺序不同的茶，其中四杯先放茶，另外四杯先放奶，次序打乱。

他这么做的原因是，他首先假设这位女士没有分辨能力（这个假设被称为原假设），如果她能很好地鉴别这 8 杯茶，则说明在原假设成立的情况下，发生了异常现象，以至于原假设是令人怀疑的。从统计上来说，如果在原假设成立的前提下发生了小概率事件，那么我们就有理由怀疑原假设的真实性——这也是假设检验的基本思想。

故事的结果不得而知，但费希尔却对此展开了深入研究。

1956 年，费希尔发表了 *6 Mathematics of a Lady Tasting Tea*，继续讨论了随机试验的重要性，以及增加样本数量和重复试验带来的益处。他在文中还讨论了实验设计中为什么“茶”和“奶”的数量应该相等。他所提出的试验设计对统计学的发展有极大的促进作用。

下面我们来详细说说到底什么是假设检验。

8.1 掀开假设检验的面纱

假设检验是数理统计学中根据一定的假设条件由样本推断总体

的一种方法。它的基本思想是小概率反证法思想。所谓小概率思想，是指小概率事件（发生的概率 $P<0.01$ 或 $P<0.05$）在一次试验中基本上不会发生。具体来说，反证法思想是先提出假设（一般称之为原假设，记为 H_0），再用适当的统计方法确定假设成立的可能性大小。若可能性小，则认为这个原假设不成立；若可能性大，则不认为原假设不成立。

在假设检验中，如果要完成一套正确的检验，除了设立最基本的原假设外，还需要建立一个与之对立的假设，称之为备择假设，记为 H_1。如果要验证所设立的原假设是否正确，则需要用从总体中抽出的样本进行检验，与此有关的理论和方法就构成假设检验的内容。常用的假设检验方法有 U 检验法、T 检验法、χ^2 检验法（卡方检验）、方差分析（F 检验法）、秩和检验等。

对于假设检验来说，有哪些需要关注的重要知识点呢？

8.1.1 原假设 VS 备择假设

原假设在统计学中又称零假设。为什么叫零假设？对于这个疑惑，理解时可能需要借助相关系数。若想判断两个事件是否相关，则可以用相关系数来测量它们之间的相关程度。之前说过，相关系数是介于–1~1 的一个数值。当相关系数是零的时候，则说这两个事件没有关系。

回到假设检验，所谓假设就是对两个事件之间关系的某种猜测。因此，如果要研究两个事件的关系，零假设的意思就是两个事件之间没有关系，而且对于零假设的内容，一般情况下希望证明其是错误的。

说完原假设（零假设），再来详细说说备择假设。前面说过，备

择假设就是原假设的对立面。此话没错，但当原假设并非单一地等于零时，它的设置可能运用了“≥”或“≤”符号，此时的备择假设也就不是简单的“≠”。当然，备择假设和原假设之间依旧是对立互补的关系，但符号之间的区别对检验结果可能会有重大影响。

当原假设中的假设关系采用等号而备择假设中的假设关系采用不等号时，将这个假设检验称作双尾检验。双尾检验需要相对较大的差异，这个差异不依赖于方向。举个例子，要比较两种药物的治疗药效如何，如果只需证明是否相等，而不管药物 A 是否比药物 B 有效，则为双尾检验。

而如果原假设中所假设的关系是有方向性的（比如大于等于或者小于等于），此时所进行的假设检验就是单尾检验。单尾检验允许在差异相对较小时就拒绝原假设，这个差异被规定了方向。同样还是两种药物的药效比较，如果要验证药物 A 比药物 B 更有效，则为单尾检验。

那么什么情况下用单尾检验，什么情况下又需要双尾检验呢？这取决于原假设的设定。一般情况下，双尾检验的要求更为严格，比单尾检验更令人信服。因为双尾检验需要有更多的证据来拒绝原假设，因此提供了更强的证据说明差异存在。但反过来，单尾检验更为敏感，即在单尾检验中相对较小的差异也可能是显著的，但是，它不能达到双尾检验的显著性要求。因此，需要针对不同的研究对象选择合适的原假设与备择假设。

在进行假设检验时，可以根据以下几个原则来建立原假设和备择假设：

（1）原假设和备择假设是一个完备事件组，而且相互对立。这样做的目的是保证两个假设中有一个必然成立，而且只能有一个成立。

（2）在建立假设的时候，先确定备择假设，然后再确定原假设，且等号“=”始终放在原假设上。

虽然是用单尾检验还是用双尾检验是由原假设决定的，但一般情况下，我们会把要反驳的观点放在原假设上，而把支持的观点放在备择假设上。事实上，我们在设立假设的时候，虽然看上去先写原假设，但潜意识里早已确立了备择假设。

但是，不管假设设置得如何周密，使用的检验方法如何精确，假设检验都是建立在一定的概率基础上的。所以，在做假设检验的时候经常会犯两类错误。

第一类错误：弃真错误

所谓弃真错误，是指原假设是真的，却拒绝了原假设。一般把犯第一类错误的概率记作α。

第二类错误：取伪错误

所谓取伪错误，是指原假设是错的，但并没有拒绝它。我们通常把犯这类错误的概率记作β。

我们用一张表格归纳一下，如表 8.1 所示。

表 8.1　假设检验的结论与后果

拒绝与否	实际情况	
	原假设为真	原假设为假
不拒绝原假设	正确	犯第二类错误
拒绝原假设	犯第一类错误	正确

可以发现，只要做出拒绝或不拒绝原假设这个决策，就会有犯错误的风险，但只能犯这两种错误中的一种。通常来说，这两种错误的概率之间存在此消彼长的关系：当α增大时，β减小；当β增大时，α减小。但在正常情况下，无论是哪种错误，我们都希望犯错的概率越小越好。

若要同时减少这两种错误发生的概率，就要增加样本量，但样本量的增加并不是无限制的。如果有过调研的经验，就会知道，样本量的确定会受到很多因素的限制，如成本、调查对象的可获取性等。

特别是在临床医学试验中，很多样本的搜集和培养都存在不可复制性。因此，针对这两类错误，必然要做出取舍。一般来说，犯第一类错误的概率往往是可以控制的，所以在进行假设检验的时候，往往先控制犯第一类错误的概率α。这个概率常被用来作为检验结论是否可靠的一个度量标志，又称显著性水平。

用统计语言来说一个检验是显著的，指的是这个样本结果并不是偶然所致的；反之，若检验发现不显著，那就说明它无法代替总体现象，只不过是偶然出现的结果。

在进一步讨论α的取值之前，先要对假设检验的统计量和它的拒绝域进行介绍。

8.1.2 检验统计量和拒绝域

既然要检验，必然需要构建一个统计量。需要说明的是，假设检验使用的数据是从总体中抽取的一组样本数据，而且默认这组样本数据是能代替总体的。在此基础上，才能构建检验统计量。

标准化检验统计量的计算公式如下：

$$\text{标准化检验统计量}=\frac{\text{点估计量}-\text{假设值}}{\text{点估计量的抽样标准差}}$$

在公式中，检验统计量反映了点估计量和假设的参数相比，它们之间到底差了几个标准差。但是这么说还是有点抽象，先来看个例子。

前面说过，在样本量足够大的情况下，样本均值的抽样分布是近似服从正态分布的，那么此时的抽样标准差就是 $\sigma/\sqrt{n}$。如果此时得到的样本均值是 $\bar{x}$，总体的均值和方差分别为 μ_0 和 σ^2（注意，这里的均值 μ_0 就是我们所假设的数值）。如果要做假设检验，此时原假设就是 $H_0:\bar{x}=\mu_0$，备择假设就是 $H_1:\bar{x}\neq\mu_0$，那么检验统计量就可以设置为 $z=\dfrac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$，其中 z 是一个随机变量。

取不同的样本值，所得的数值也是不同的。这类似于将普通正态分布转换为标准正态分布的过程。的确，这个检验量经过对样本数据的标准化后服从的正是标准正态分布。

有了统计量，接下来就是代入数据求出数值。但是，有了数值并没什么用，关键是根据这个数值怎么判断拒不拒绝。

有了一个数值，就要想办法制定一个划分规则来判断什么数值出现时该拒绝原假设，什么数值出现时不能拒绝原假设。我们称这个划分规则为拒绝域。

所谓拒绝域，具体来说就是由显著水平α围起来的区域。一般而言，α的取值按照检验精度要求不一可选择 0.01、0.05 和 0.1 三个数值。不过由于 0.01 这个数据要求比较高，常用于医学检验，因此，

在大多数情况下，选择 0.05 就够了。当然，也可以根据实际情况选择其他的概率数值（只要在 0～1 即可）。

知道了α的取值，又怎么来求它围起来的区域呢？这就需要通过α的值来查找相应的分布表，以此来获得区域两端的临界值。

下面以正态分布的拒绝域图示为例，来帮助大家更好地理解“拒绝域”这个概念，如图 8.1 所示。

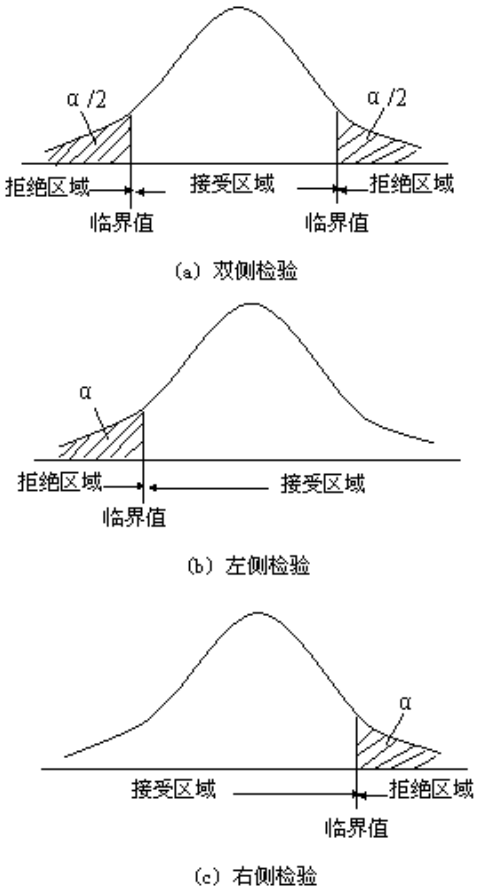


图 8.1　正态分布拒绝域图示

图 8.1 中的双侧检验和单侧检验就是上文所说的双尾检验和单尾检验，在实际运用中需根据原假设来进行选择。比如，原假设是 $H_0: \bar{x} = \mu_0$，那么在寻找拒绝域的时候就要参考图 8.1（a）。

从图 8.1 中不难看出，当样本量固定时，拒绝域取决于α的大小。α的数值越大，拒绝域的面积就越大（斜线面积=拒绝域面积=α）。

8.1.3 *P* 值

P 值是什么？它就是一个概率值，一个成名于费希尔之手的概率值。先来简单说说什么是 *P* 值。

P 值就是当原假设为真时，比所得到的样本观察结果更极端的结果会出现的概率。如果 *P* 值很小，则说明这个极端情况的发生概率很小；而如果一旦出现了，根据小概率原理，我们就有理由拒绝原假设。*P* 值越小，我们拒绝原假设的理由越充分。总之，*P* 值越小，表明结果越显著。但是检验的结果究竟是“显著的”、“中度显著的”还是“高度显著的”，则需要根据 *P* 值的大小和实际问题来解决。

通常，只要 *P* 值小于显著性水平α，就认为在既定原假设为真的情况下出现的结果如此极端，以至于我们不再相信原假设本身。一句话，我们的判定法则是：如果 *P* 值小于显著性水平α，则拒绝原假设；否则只能不拒绝原假设。

当时，费希尔认为假设检验是一种程序，研究人员依照这一程序可以对某一总体参数形成一种判断。也就是说，他认为假设检验是数据分析的一种形式，是人们在研究中加入的主观信息。不过这一观点遭到了奈曼和皮尔逊的反对。我们来看看费希尔是如何进行假设检验的：

（1）假定某一参数的取值。

（2）选择一个检验统计量，该统计量的分布在假定的参数取值为真时应该是完全已知的。

（3）从研究总体中抽取一个随机样本。

（4）计算检验统计量的值。

（5）计算概率 P 值或者观测的显著水平，即在假设为真的前提下，检验统计量大于或等于实际观测值的概率。

- $P<0.01$，说明是较强的判定结果，拒绝假定的参数取值。
- $0.01<P<0.05$，说明是较弱的判定结果，拒接假定的参数取值。
- $P>0.05$，说明结果更倾向于接受假定的参数取值。

这个程序步骤其实和我们之前所讲的假设检验步骤基本一致，只是在第一步的时候未提到备择假设，以及在最后做判断的时候一个用简单的 P 值，另一个则是参考拒绝域的临界值，但其本质是一样的。

不过，在实际运用中，还需要注意以下几点：

（1）P 的意义不表示两组差别的大小，P 反映的是两组差别有无统计学意义。

（2）当 $P>\alpha$ 时，差异无显著意义，根据统计学原理可知，不能否认原假设，但并不认为原假设肯定成立。

（3）统计学主要用三种 α 值来与 P 值作比较（0.1；0.05；0.01），也可以计算出确切的 P 值，也有人用 $P<0.001$，至于选择哪个，要看

检验的应用领域。

（4）显著性检验只是统计结论，判断差别还要根据专业知识。

其实，在生活和工作中到处都充斥着假设检验的身影，比如我们目前经常用到的各类 APP，时常看到工程师不遗余力地更新版本，除了对 APP 功能的升级外，有时也会对软件的操作界面进行优化。有一个广泛应用的测试——A/B 测试，就运用了假设检验的思想。比如在 APP 的设计上，检验者制定了软件的两个操作界面，可能在标题字体、背景颜色、措辞等方面有所不同，然后将这两个界面以随机的方式同时推送给所有用户。其中一部分用户使用 A 方案，另一部分用户使用 B 方案，检验者记录下用户的使用情况，然后拟定检验量来判断哪个方案更符合用户的审美。当然，假设检验在药物的药效检验中更为普遍，几乎所有药物都会经过各类假设检验。这种检验可能不仅仅针对两种药物，也可能是多种药物间的药效比较。

多种药物的药效如何比较？难道是逐一比较？还记得我们之前说过的几类假设检验经常使用的检验方法吗？下面会略作介绍。

8.2 几种常用假设检验简介

1．对均值的检验

对于均值而言，有 T 检验（U 检验）和方差分析两种方法。为什么说是两种方法？因为 U 检验和 T 检验是基本一致的，只是使用的条件略有不同（当样本含量 n 较大时，样本观测值符合正态分布，可用 U 检验进行分析；当样本含量 n 较小时，若观测值符合正态分布，则用 T 检验）。为了方便行文，我们以 T 检验作为代表，来看看它和方差分析之间的联系与区别。

T 检验只能用于两样本均数及样本均数与总体均数之间的比较。根据研究设计，T 检验可以分为三种形式。

（1）单个样本的 T 检验：又称单样本均数 T 检验，适用于样本均数与已知总体均数的比较，目的是检验样本均数所代表的总体均数是否与已知总体均数有差别。

（2）配对样本均数 T 检验（非独立两样本均数 T 检验）：适用于配对设计的数据均数的比较，目的是检验两相关样本均数所代表的未知总体均数是否有差别。

（3）两个独立样本均数 T 检验：适用于完全随机设计的两样本均数的比较，目的是检验两样本所来自总体的均数是否相等。

方差分析主要是用于两样本及以上样本之间的比较，又被称为“变异数分析”或“F 检验”。方差分析的基本思想是：通过分析研究不同来源的变异对总变异的贡献大小，从而确定可控因素对研究结果影响力的大小。方差分析的一个基本逻辑是：按照可分解性的原则，总变异可以被分解成组间变异和组内变异。组间变异是指由于不同的试验处理而造成的各组之间的变异；组内变异是指组内各被试变量的差异范围所呈现的变异。组间变异对组内变异的比值越大（这个比值也就是检验统计量，服从 F 分布），各组均值的差异就越大，以此来判断各组间均值差异是否显著。它有 4 个主要用途：

（1）多样本均值差别的显著性检验。

（2）分离各有关因素并估计其对总变异的作用。

（3）分析因素间的交互作用。

（4）方差齐性检验。

T 检验和方差分析的联系：

（1）二者都要求比较的资料服从正态分布。

（2）两样本均数的比较及方差分析均要求比较组有相同的总体方差。

（3）配对组比较的方差分析是配对比较 T 检验的推广，成组设计多个样本均数比较的方差分析是两样本均数比较 T 检验的推广。

（4）对于两个样本之间的比较，方差分析和 T 检验的效果是相同的。

T 检验和方差分析的区别：T 检验只能用于两样本均数的比较，而方差分析可以用于多样本之间的比较。

2．对频率/比值的检验

除了对均值进行检验外，有时还会针对比率进行检验，这时就要用到卡方检验。卡方检验既可以用来检验两个比率，也可以用来判断多个比率之间是否存在差异。

卡方检验除了可以检验比率差异外，还有其他一些用途：

（1）检验某个连续变量的分布是否与某种理论分布相一致。如变量是否符合正态分布、是否服从均匀分布、是否服从泊松分布等。

（2）检验某两个分类变量是否相互独立。比如吸烟与否与是否患有呼吸道疾病之间的关联。

（3）检验某两种方法的结果是否一致。比如采用两种诊断方法对同一批人进行诊断，其诊断结果是否一致；采用两种方法对客户进行价值类别预测，预测结果是否一致等。

除了均值和比率检验外，很多地方都会用到假设检验，也可以根据实际情况灵活运用各类检验方法。但是，如何才能快速、有效、准确地得出假设检验的分析结果？那些结果又该怎么解读呢？

8.3 手把手教你做检验

要进行数据分析，首先要有数据。我们使用的数据分为试验组 A 和实验组 B，两组数据各有 206 个观测值。使用的软件是 SPSS 2.0 版本。

下面正式步入实证操作。

导入数据的过程不再详述，直接从两样本的 T 检验开始。首先来看单样本的 T 检验，目的是检验数据均值是否与指定的均值相同（通常也就是样本均值是否等于总体均值的检验）。以 A 组数据来进行演示：

首先通过“分析”→“比较均值”→“单样本 T 检验”菜单命令进入分析界面，如图 8.2 所示。

在“单样本 T 检验”对话框中，将 A 组数据选作检验变量，检验值设置为 20000，当然也可以输入其他检验值。一旦输入了检验值，就确定了原假设 H_0：A 组均值=20000，以及备择假设 H_1：A 组均值≠20000。接着单击“选项”按钮，在“置信区间百分比”文本框中输入 95%，依次单击“继续”和“确定”按钮，如图 8.3 所示，就可以得到检验结果，如表 8.2 所示。

图 8.2　SPSS 假设检验操作界面

图 8.3　单样本 T 检验操作选项

表 8.2　单样本统计量

	N	均值	标准差	均值的标准误
A 组	206	215991.05	1128741.686	78643.160

在表 8.2 中，可以看到对 A 组数据的描述性统计，观测值个数为 206 个，均值为 215991.05。我们当时输入的检验值是 20000，这个样本均值和 20000 有差异吗？继续看看软件输出的第二张表格，如表 8.3 所示。

表 8.3 单样本 T 检验结果

	检验值 = 20000					
	t	*df*	Sig.(双侧)	均值差值	差分的 95%置信区间	
					下限	上限
A 组	**2.492**	**205**	**0.013**	195991.049	40937.92	351044.18

在表 8.3 中，我们看到 T 检验统计量的值为 2.492，检验的 *P* 值为 0.013。我们选择的显著性水平是 0.05，这意味着通过 *P* 值与α的比较就能直接得出结论：拒绝原假设。也就是说，215991.05 和 20000 是有区别的。

我们会了单样本的 T 检验，再来看看两个配对样本的 T 检验又该怎么做。

仍然使用这两组数据 A 和 B，假设它们是配对的（所谓配对，是指两组数据观测值数量一致，而且 A 和 B 之间的观测值顺序不可改变）。此时，检验结果会如何呢？

这次选择“分析”→“比较均值”→“配对样本 T 检验”菜单命令进入分析界面，然后在弹出的对话框中将 A 组和 B 组分别选入第一组配对中，同样在“选项”按钮中保持默认设置，如图 8.4 所示，单击“确定”按钮后得到如表 8.4 所示的结果。

图 8.4 配对样本 T 检验操作界面

从表 8.4 中可以看到，A 组和 B 组的观测值均为 206，没有样本损失。但两组的均值却不太一样：A 组依旧是 215991.05，而 B 组的均值则为 156941.17。从直观上看来这两组数据的均值明显不相等。那此时的原假设是什么呢？应该是 H_0：A 组均值=B 组均值，以及备择假设 H_1：A 组均值≠B 组均值。那么统计检验结果是不是如我们料想的那样呢？再来看表 8.5。

表 8.4　配对样本统计量

		均值	*N*	标准差	均值的标准误
对 1	A 组	215991.05	206	1128741.686	78643.160
	B 组	156941.17	206	714932.513	49811.709

表 8.5　配对样本 T 检验结果

		成对差分					*t*	*df*	Sig. (双侧)
		均值	标准差	均值的标准误	差分的 95%置信区间				
					下限	上限			
对 1	A 组 –B 组	59049.88	461577.44	32159.624	–4356.1	122455.91	**1.836**	**205**	**0.068**

在表 8.5 中可以看到，T 检验统计量的值为 1.836，对应在 0.05 的显著性水平下，*P* 值为 0.068，大于 0.05。也就是说，我们并不能拒绝认为两组均值无差别的原假设。当然，如果我们放宽检验精度，将显著性水平值设置为 0.1，依旧可以拒绝原假设。这就是 *P* 值的作用。

事实上，在日常生活和工作中，我们对数据的假设检验并没有这么简单，往往涉及分组实验的设计、多因素的分析比较等内容，现如今有了统计软件的帮助，让分析工作减轻了许多。不过每一种分析如何进行，由于市面上的软件品种繁多，方法和参数的设置也各不相同，所以不在此赘述。不过，笔者有个疑问，那就是我们所进行的假设检

验，前提都是可以获得检验使用的数据，这些数据往往是经过随机试验设计而获得的。如果我们无法获得更多的样本，甚至无法得到样本，那么该怎么办？

别着急，你马上就会得到答案。

☆本章知识拓展

本章我们对各种常用的假设检验方法进行了介绍，下面将各个常用检验方法的检验假设、检验统计量及对应的拒绝域整理如下，如表 8.6 和表 8.7 所示。

表 8.6 单个正态总体均值、方差的假设检验

检验法	条件	H_0	H_1	检验统计量	拒绝域
Z 检验	σ已知	$\mu \leqslant \mu_0$	$\mu > \mu_0$	$z=\dfrac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$	$\{z \geqslant z_{1-\alpha}\}$
		$\mu \geqslant \mu_0$	$\mu < \mu_0$		$\{z \leqslant z_{\alpha}\}$
		$\mu = \mu_0$	$\mu \neq \mu_0$		$\left\{\lvert z\rvert \geqslant z_{1-\frac{\alpha}{2}}\right\}$
Z 检验	σ未知 （大样本）	$\mu \leqslant \mu_0$	$\mu > \mu_0$	$z=\dfrac{\overline{X}-\mu_0}{s/\sqrt{n}}$	$\{z \geqslant z_{1-\alpha}\}$
		$\mu \geqslant \mu_0$	$\mu < \mu_0$		$\{z \leqslant z_{\alpha}\}$
		$\mu = \mu_0$	$\mu \neq \mu_0$		$\left\{\lvert z\rvert \geqslant z_{1-\frac{\alpha}{2}}\right\}$
T 检验	σ未知 （小样本）	$\mu \leqslant \mu_0$	$\mu > \mu_0$	$t=\dfrac{\overline{X}-\mu_0}{s/\sqrt{n}}$	$\{t \geqslant t_{1-\alpha}(n-1)\}$
		$\mu \geqslant \mu_0$	$\mu < \mu_0$		$\{t \leqslant t_{\alpha}(n-1)\}$
		$\mu = \mu_0$	$\mu \neq \mu_0$		$\left\{\lvert t\rvert \geqslant t_{1-\frac{\alpha}{2}}(n-1)\right\}$
χ^2 检验	μ未知	$\sigma^2 \leqslant \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\chi^2=\dfrac{(n-1)s^2}{\sigma_0^2}$	$\{\chi^2 \geqslant \chi^2_{1-\alpha}(n-1)\}$
		$\sigma^2 \geqslant \sigma_0^2$	$\sigma^2 < \sigma_0^2$		$\{\chi^2 \leqslant \chi^2_{\alpha}(n-1)\}$
		$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$		$\left\{\lvert\chi^2\rvert \geqslant \chi^2_{1-\frac{\alpha}{2}}(n-1)\right\}$ 或 $\left\{\lvert\chi^2\rvert \leqslant \chi^2_{\frac{\alpha}{2}}(n-1)\right\}$

表 8.7　两正态总体均值、方差的假设检验

检验法	条件	H_0	H_1	检验统计量	拒绝域
Z 检验	σ_1,σ_2 已知	$\mu_1 \leqslant \mu_2$	$\mu_1 > \mu_2$	$z=\dfrac{\bar{x}-\bar{y}}{\sqrt{\dfrac{\sigma_1^2}{n}+\dfrac{\sigma_2^2}{m}}}$	$\{z \geqslant z_{1-\alpha}\}$
		$\mu_1 \geqslant \mu_2$	$\mu_1 < \mu_2$		$\{z \leqslant z_{\alpha}\}$
		$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$		$\left\{\lvert z\rvert \geqslant z_{1-\frac{\alpha}{2}}\right\}$
T 检验	$\sigma_1=\sigma_2$ 未知	$\mu_1 \leqslant \mu_2$	$\mu_1 > \mu_2$	$t=\dfrac{\bar{x}-\bar{y}}{s_w\sqrt{\dfrac{1}{n}+\dfrac{1}{m}}}$	$\{t \geqslant t_{1-\alpha}(n+m-2)\}$
		$\mu_1 \geqslant \mu_2$	$\mu_1 < \mu_2$		$\{t \leqslant t_{\alpha}(n+m-2)\}$
		$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$		$\left\{\lvert t\rvert \geqslant t_{1-\frac{\alpha}{2}}(n+m-2)\right\}$
近似Z检验	$\sigma_1=\sigma_2$ 未知（大样本）	$\mu_1 \leqslant \mu_2$	$\mu_1 > \mu_2$	$z=\dfrac{\bar{x}-\bar{y}}{\sqrt{\dfrac{\sigma_x^2}{n}+\dfrac{\sigma_y^2}{m}}}$	$\{z \geqslant z_{1-\alpha}\}$
		$\mu_1 \geqslant \mu_2$	$\mu_1 < \mu_2$		$\{z \leqslant z_{\alpha}\}$
		$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$		$\left\{\lvert z\rvert \geqslant z_{1-\frac{\alpha}{2}}\right\}$

第 9 章

回归分析——科学研究的“万金油”

在上一章的结尾处留了一个疑问：虽然随机试验可以帮助我们发现很多数据间的关系，但在一些触及“底线”的情形下，当我们无法通过随机试验进行数据分析时，又该怎么办呢？

本章就来回答这个问题，这就需要用到回归分析。

“回归”一词的诞生过程如下：

1855 年，高尔顿和当时还是他学生的老皮尔逊两人通过观察 1978 对父子的身高数据，发现这些数据如果用散点图画出来，则近乎是一条直线。也就是说，从 1978 对父子的身高数据中发现这样一个趋势：当父亲的身高增加时，儿子的身高也倾向于增加。但这不是最重要的，吸引高尔顿的是另一个有趣的现象：当父亲的身高高于平均身高时，其儿子的身高比其自身身高更高的概率要小于比其更矮的概率；而那些低于平均身高的父亲，其儿子的身高比他更矮的概率要

小于比他更高的概率。于是，一个推论在高尔顿的脑海中浮现：难道这两种身高的父亲，其儿子的身高有向他们父辈的平均身高回归的趋势？

对于如此惊人的发现，高尔顿左思右想，最终认识到这必定是真实的，而且在进行所有观察之前是可以预言的。高尔顿的思想过程为：假设不发生这种向平均值的回归，那么从平均意义上看，高身高父亲的儿子将与他们的父亲一样高。在这种情况下，一些儿子的身高必须高于他们的父亲，以抵消身高比父亲矮小者的影响，使平均值不变。高身高儿子这一代人的儿子也将如此，那么会有一些孙子身高更高。这个过程将一代一代地延续下去。同样，将会有一部分儿子身高比他们父亲矮小，而且有一部分孙子将更加矮小。如此下去，不用多少代，人类种族就将发生两极分化。但是这种情形并没有发生，人类的身高在平均意义上仍趋向于保持稳定。也就是说，只有当非常高的父亲其儿子的平均身高变矮，而非常矮的父亲其儿子的平均身高变高时，才能出现这种稳定，高尔顿称其为“回归”。

9.1 探寻“回归”的本质

首先来看看回归分析的一个现代思想：它早已不是简单的回归均值的问题，而是讨论一个模型中解释变量（自变量）和被解释变量（因变量）之间的关系。在之前相关分析的篇章里，我们说过，变量间未必都是确定的函数关系，这时回归分析就派上了用场。

对于数据而言，要运用回归分析，也是有一定的假设要求的，如下：

（1）随机误差项是一个期望值或平均值为 0 的随机变量。

（2）对于解释变量的所有观测值，随机误差项有相同的方差。

（3）随机误差项彼此不相关。

（4）解释变量是确定性变量，不是随机变量，与随机误差项彼此之间相互独立。

（5）解释变量之间不存在精确的（完全的）线性关系，即解释变量的样本观测值矩阵是满秩矩阵。

（6）随机误差项服从正态分布。

假设虽多，但最为关键的是两个概念：解释变量和随机误差项。首先来看解释变量。对于解释变量，回归分析要求的不多：如果是确定变量，且有多个解释变量之间彼此不相关，则“回归”能派上用场；否则回归分析并不适用。比如，想要研究高学历的父母是不是容易生出高学历的孩子。这个问题涉及的因素比较多。我们可以选择孩子的学历作为被解释变量，父母的学历作为解释变量。其中，父母的学历是既定的，所以它符合第（4）条假定的前半部分，但其实和孩子学历高低有关的变量还有很多，如家庭的经济环境、父母对于教育的重视程度等，而这些也可以作为解释变量纳入研究模型中。不过有些变量在选择上需要注意，如父母的智商就不太适合纳入，因为智商和学历有比较紧密的关系。同样，学历高低和收入高低也有一定的关系，当这些变量一同放入回归模型时，就违反了第（5）条假设。

关于解释变量，还有一个问题，即第（4）条假设的后半句：解

释变量与随机误差项彼此之间相互独立。这怎么理解？我们再回到教育问题上。比如，想要研究中学生的学习成绩和逃课率之间的关系，可以简单地将成绩作为被解释变量，逃课率作为解释变量，建立一个回归方程。此时有哪些因素可以归入随机误差项呢？比如天气、性别等，显然在中学生范围内，这些因素和逃课率之间并没什么关系。

接下来看看随机误差项。随机误差项在上文已经略有涉及，此处要说的重点是随机误差项的零均值与方差和不相关。为什么要让随机误差项是零均值同方差呢？因为只有这样，才能让参数估计是无偏和有效的。抛开参数估计，来说说随机误差项的零均值。既然有随机误差，那么它就可能有大有小，但最终它的均值需要为 0；同样，还需要保证随机误差对变量的影响程度是相等的。比如，进行一项农业研究，在几块土地上施以不同的肥料浓度来观测施肥浓度对粮食产量的影响，其中随机变量就可能是天气、湿度等因素，但是对于实验土地来说，这些因素的影响是均等的。而对于随机变量彼此不相关，也就是说不存在序列相关这个假设条件，则可以这样理解：随机变量是不随时间变动而变动的。仍然研究施肥浓度对粮食产量的影响问题，也许某天刮了一场台风、下了一场瓢泼大雨，当季的粮食产量受到严重影响，这是一个随机影响，但这个影响是一次性的。这就是随机变量不相关的表现。

既然对数据有那么多假设，那么“回归分析”又能帮助我们解决什么问题呢？除了可以帮助存在相关关系的变量找出数学表达式外，还能借助这个表达式起到预测的功能，最重要的是不仅能预测，还能控制精度。

那么，回归分析有哪几种模型可供选择呢？比如，按照解释变量的个数，可以将回归分为一元回归和多元回归；按照数据的分布情况，又可以将回归分为线性回归和非线性回归；根据不同数据的特点，该模型还可以细分为 Linear Regression、Logistic Regression、Probit Regression 等。

9.2 释放“回归”的超能力

要让“回归”发挥实力，首先要知道进行回归分析的步骤。

（1）确定回归方程中的解释变量和被解释变量。

这一步至关重要。对于熟悉解应用题的学生来说，要分清楚解释变量和被解释变量很简单，看谁是 X 谁是 Y 就可以了。但在现实分析中，解释变量和被解释变量都是需要我们自己来判断的。比如，想要研究广告投入费用与销售业绩之间的关系式，你能立即判断出谁是 X 谁是 Y 吗？通常情况下，把要研究的对象作为被解释变量，把其他能找到的辅助研究该对象的变量列为解释变量。比如在这个例子中，我们关注的其实是销售业绩，那么就把销售业绩作为被解释变量，把广告投入作为解释变量来辅助研究。

（2）确定回归模型。

根据函数拟合方式，通过观察散点图确定通过哪种数学模型来描述回归线。如果被解释变量和解释变量之间存在线性关系，则进行线性回归分析，建立线性回归模型；如果被解释变量和解释变量之间存在非线性关系，则进行非线性回归分析，建立非线性回归模型。

（3）建立回归方程。

根据收集到的样本数据及确定的回归模型，在一定的统计拟合准则下估计出模型中的各个参数，得到一个确定的回归方程。这里的拟合准则、估计方法其实主要就是之前讲到的最小二乘法和最大似然估计法。

（4）对回归方程进行各种检验。

由于回归方程是在样本数据的基础上得到的，它是否真实地反映了事物总体间的统计关系，以及能否用于预测等，都需要进行检验。常用的检验方法是假设检验。

（5）利用回归方程进行预测。

这一步并不是必需的，因为并不是所有的回归分析都需要进行预测，有时候我们要关注的重点可能仅仅是找出变量间的关系。

至于怎么做回归分析，下面通过一个小例子来说明。

假如要研究儿童的体重与身高之间的关系，该怎么做？在不知道回归分析这门技术之前，我们也许会选择先用描述性统计对数据做一番统计描述，然后做一个相关分析以证明二者是相关的。现在有了“回归分析”，那么我们的研究可以深入，将二者的关系量化。

仍然使用 SPSS 软件来帮助我们进行分析。在 SPSS 中，回归分析位于“分析”模块中，而在选择之前，首先要判断我们所使用的数据是适合使用线性回归还是非线性回归。先来画一幅散点图，如图 9.1 所示。

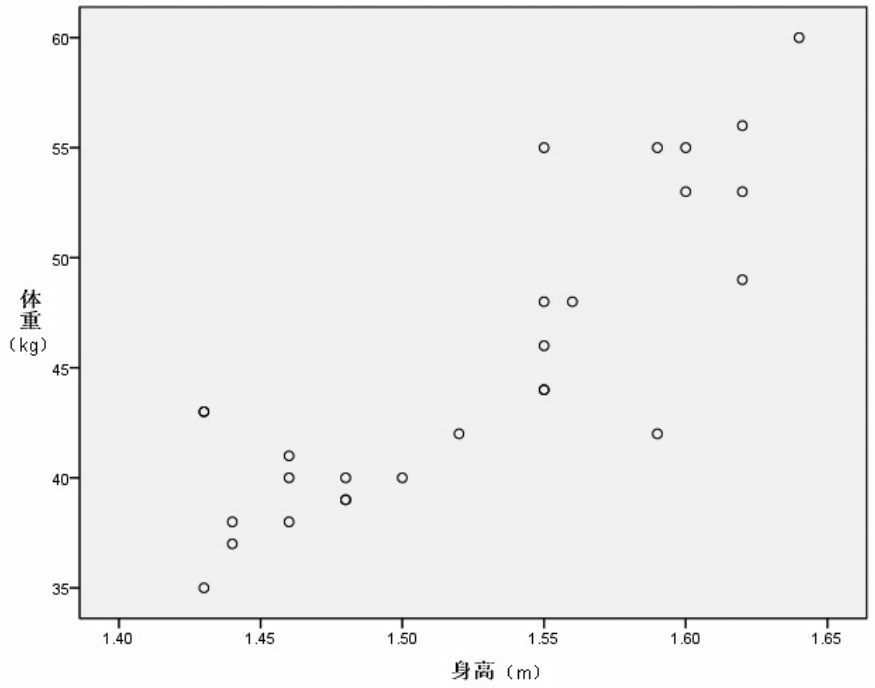


图 9.1 体重与身高散点图

从散点图中可以大致判断身高和体重是呈线性关系的，所以可以考虑用线性模型来拟合直线。不过仅有散点图还不够，还需要变量之间有线性相关关系。我们来看看二者的相关关系如何，如表 9.1 所示。

表 9.1 身高与体重相关系数表

		身高	体重
身高	Pearson 相关性	1	0.849**
	显著性（双侧）		0.000
体重	Pearson 相关性	0.849**	1
	显著性（双侧）	0.000	

**在.01 水平（双侧）上显著相关。

从表 9.1 中可以直观地看到，身高和体重之间的相关系数达到 0.849，相关度比较高。再结合散点图，就可以进行线性回归分析了。

在“分析”模块中勾选“回归”→“线性”，进入线性回归模型的设置对话框，如图 9.2 所示。

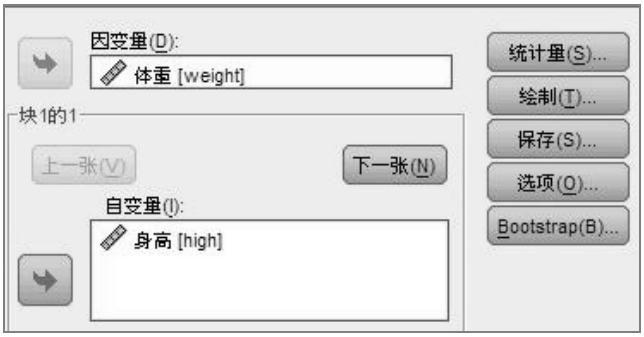


图 9.2　线性回归设置对话框

因为我们要研究的是体重和身高的关系，而且通常身高会影响体重，所以在模型设置对话框中将因变量（被解释变量）选为体重，自变量（解释变量）选为身高，其他保持默认选择即可，然后单击“确定”按钮。

现在我们会得到哪些结果呢？

首先来看看回归模型的系数检验表，如表 9.2 所示。从表中可以看到，身高的系数为 85.129，对应的 t 值是 8.030。那么这个检验统计量是否落在拒绝域呢？我们看 t 值右边的 Sig 值，这其实就是之前我们所说的 P 值，目前它非常接近于 0，小于显著性水平 0.05，所以我们认为，对于身高这个变量来说，系数 85.129 是显著有效的。同样，我们也可以判断出常量也是有效的。至此，可以将身高和体重之间的关系描述为：

$$体重=85.129\times身高-84.605+\varepsilon$$

虽然我们拥有数学表达式，但并不能解决所有的疑惑，比如有人

会问，即便 85.129 这个系数有效，那又量化了什么？

表 9.2 系数检验表

模型		非标准化系数		*t*	Sig.
		系数	标准误差		
1	（常量）	−84.605	16.193	−5.225	0.000
	身高	85.129	10.601	8.030	0.000

对于回归系数的理解，笔者认为，在多元回归模型中，在其他变量不变的情况下，每个解释变量的回归系数代表每提升（或降低）一个单位，相对应的被解释变量平均变动该系数个单位。如果是一元回归，那么回归系数就和斜率的概念相一致。

再来看表 9.3。

表 9.3 模型拟合情况

模型	*R*	*R* 方	调整 *R* 方	标准估计的误差
1	0.849[a]	0.721	0.709	3.752

这个表说的是什么？R 方又是什么？R 方是可决系数，也被称为拟合优度，它是一个取值在[0,1]的数。在线性模型中，该值越接近 1，则说明模型与原始数据越贴合。在这个例子中，*R* 方为 0.721，调整的 *R* 方也达到 0.709，可见，模型的拟合效果可以接受。为何还有一个调整的 *R* 方？

在建立多元回归模型时，如果增加解释变量，则可决系数也会随之逐渐增大，当解释变量足够多时，我们会错误地认为模型拟合良好，而实际情况却并非如此。于是我们就会考虑对 *R* 方进行调整，称之为调整后的 *R* 方。而且在多元回归中，调整的 *R* 方会得到更多关注。

最后来看表 9.4。

表 9.4　模型整体检验

模型		平方和	*df*	均方	*F*	Sig.
1	回归	907.698	1	907.698	64.480	0.000[b]
	残差	351.931	25	14.077		
	总计	1259.630	26			

表 9.4 的输出结果就是模型总体的 F 检验。从 Sig 值可以看出，我们是可以拒绝原假设的。这里的原假设是什么？在回归分析中，如果解释变量为 $x_1, x_2, \cdots, x_n$，那么系数 a_i 的 T 检验原假设是 $a_i = 0$；而 F 检验的原假设则是 $a_1 = a_2 = \cdots = a_i = 0$。如果我们做的是一元回归分析，那么此时的 T 检验和 F 检验是等价的；但在多元回归里，F 检验就很重要，它的意义在于检验模型是否整体有效。从表 9.4 中可以看出，我们的模型是有效的。

9.3　规避“回归”的误区（伪回归问题）

在回归分析中，如果稍有失误，就可能会陷入误区。

误区 1：样本量过小——你的样本有代表性吗

在上一节的回归分析中，样本量是多少？在整个回归模型建立的过程中，笔者都未对此加以说明。事实上，笔者的这批数据是 27 个儿童身高和体重的样本，用 27 个儿童来代替整体并不可靠。随之而来的疑问就是：样本能代表总体吗？是的，无论是假设检验还是回归分析，我们都希望透过样本来发现总体规律。

我们知道，某些临床实验会采用小样本（或者对于一些罕见病来说，只能获得小样本），而这就增加了随机偏离的数据在统计中起到的作用，使得研究结果有偏，而这只是一种客观上的小样本。另一种则是有意无意地缩减样本。比如观察某类药物的摄取量对该疾病的治

疗效果，如果选择观测时间为三个月，通过回归分析可能得到的是随着药物摄取量的增加，疾病治疗效果越为显著。如果不进行更长时间的监测，那么或许就无法发现当药物摄取量达到某一值后对疾病治疗已无明显作用，甚至继续增加药物摄取量将会导致其他不良症状，这时候，“小样本”就失效了。

为此，我们需要尽可能地获得大样本（一般 N 大于 30），或者保证数据的正态性，这样才能得到真正有价值的结论。

误区 2：未对回归分析的前提假设进行检验

虽然我们已经知道了回归分析的假设条件，但对于有些假设，如果不事先建立模型，是无法对它做出检验的，如随机误差项是否不存在自相关、随机误差项是否是同方差等。所以在完成了上文所说的主要检验后，还需要对模型的随机误差项做一系列检验，包括误差项的正态性检验——QQ 图/PP 图、误差项的异方差检验——White 检验、误差项的自相关检验——DW/LM 检验等。

同样，针对解释变量，我们也要对其是否具有完全共线性进行检验。共线性检验其实可以从相关系数 T 检验中事先获得一些信息。比如，在做多元回归分析时发现，如果将每个解释变量分别与被解释变量做一元回归，则回归系数都是显著有效的；而放在一起做多元回归时，却总有几个变量的 T 检验无法拒绝原假设，此时就意味着解释变量极有可能存在严重的共线性问题。

当完全共线性发生时，会对回归分析造成以下影响：

（1）完全共线性下参数估计量不存在。

（2）参数估计量的经济含义不合理。

（3）变量的显著性检验失去意义，可能将重要的解释变量排除在模型之外。

（4）模型的预测功能失效：变大的方差容易使区间预测的“区间”变大，使预测失去意义。

那么，可以通过什么方法进行检验和规避呢？常用的检验方法是通过 VIF（方差膨胀因子）是否大于 10 来进行简单判断；如果需要规避修正的话，也有很多方法，比如可以选择用逐步回归、岭回归、主成分法提取变量等来代替普通的线性回归。

误区 3：“伪回归”——真真假假分不清

“伪回归”中的“伪”指的是虚假相关关系。我们在衡量两组数据是否有相关关系的时候，无非就是将它们放在一起画一张散点图，计算相关系数，然后得出是否有线性相关关系。其实我们并没有告知软件这两组数据的定义。换句话说，如果我们将树木高度与中国 GDP 放在一起进行相关分析，那么也能得到一个很高的相关系数，但事实是，二者之间没有任何关系。

但在有些时候，“伪回归”的存在是由于思维的定式和逻辑的漏洞引起的，并非故意而为，所以“伪回归”不仅考量我们的科学道德，而且还考量我们的逻辑思维能力。

当然，并非避开了这三大误区我们就能大步向前了，在回归分析的过程中，还会出现诸多小误区，比如遗漏了重要的解释变量，从而造成回归结果的严重偏差。举例来说，当我们研究与疾病相关的影响因素时，性别、年龄这两个变量就不能遗漏。因为它们对于人体的各类疾病都存在大大小小的影响，一旦遗漏，就会造成其他解释变量的

回归系数出现偏差。如果只是系数的数值大小偏差那么还可以挽回，但如果直接造成系数正负颠倒，就会成为致命的错误。

所以，“回归”这个武器用得好，能让分析工作事半功倍；一旦失误，就会成为伤人的利器。因此，我们在进行回归分析的过程中需认真加以识别、确认、检验、修正。

☆本章知识拓展

首先来明确一个基本前提，那就是回归分析与相关分析的联系和区别。

区别：

（1）相关分析中的两个变量的地位是相等的，而回归分析中的变量则需要分为解释变量和被解释变量。

（2）相关分析中的两个变量都是随机变量，而回归分析中只有被解释变量是随机变量。

（3）相关分析适用于判定相关程度和方向，而回归分析则可以进一步地进行模型预测和控制。

联系：

（1）相关分析是回归分析的基础和前提，没有相关则无法进行回归。而且对于线性模型来说，相关程度越高，回归效果越好。

（2）相关分析和回归分析的理论方法具有一致性，一般来说，相关系数和回归系数的方向一致，可以互相推算。

（3）回归分析是相关分析的继续和深化。

按照不同的划分规则，回归也可以分为几个类别。下面介绍几个典型的回归模型。

（1）Logistic 回归：它是除线性回归外应用范围最广的。Logistic 回归与线性回归不同，它要求被解释变量必须是分类变量，不可能是连续变量。分类变量既可以是二分类；也可以是多分类，多分类中既可以是有序，也可以是无序。Logistic 回归有个近邻叫 Probit 回归，二者不仅函数模式十分接近，而且分析结果也类似。不过 Probit 回归的实际含义不如 Logistic 回归容易理解。

（2）cox 回归：cox 回归是回归家族里的一个另类，因为 cox 回归的被解释变量有些特殊：它的被解释变量必须同时有两个，一个代表状态，所以是分类变量；另一个代表时间，所以是连续变量。只有同时具有这两个变量，才能使用 cox 回归分析。cox 回归主要用于生存资料的分析。

（3）主成分回归：主成分回归其实是将主成分分析与线性回归结合在一起。所谓的主成分分析就是把多个具有高度相关的变量所包含的信息用一个或两三个变量来表示，我们称这个变量为主成分。

（4）岭回归：又称脊回归，由于模型的解与正则化参数 λ 之间的图像类似于山脊，因此得名。岭回归作为修正变量完全共线性的方法，其思路为：既然线性模型在解释变量完全共线的时候估计值会不稳定，那么岭回归在最小二乘估计里加个 k 值，改变它的估计值，使估计结果变稳定。至于 k 值的确定，可以先选很多个 k 值，然后作出岭迹图，看看这个图在 k 取哪个值的时候较为稳定，选取该 k 值即可。

（5）偏最小二乘回归：该回归可以用于解决解释变量之间高度相关的问题，其优势是可以用于样本量很少的情形。它的原理其实跟主成分回归类似，即用被解释变量和解释变量的综合变量来进行分析，所以它也可以用于多个解释变量的回归。这么说来，偏最小二乘法集主成分分析、典型相关分析和多元线性回归分析三种分析方法的优点于一身，成为分析领域的“新贵”。

第 10 章

物以类聚，人以群分

“类别”在现今是一个词，但若拆分开，它却是完全相悖的两个词：类，有种类、相似之意；而别，则意为区别。一个求同、一个存异，在这里，却能完美地演绎聚类分析和判别分析。下面先从求同开始，说说聚类分析。

10.1 分久必合——聚类分析

要说聚类分析，还要追溯到分类学。在很久以前，分类学就立志用经验和专业知识来为人类实现对事物的分类，但在当时并没有科学理论和数学工具来帮助人们进行定量的分类。但是时代的车轮总是向前的，随着人类科学技术的发展，对分类的要求也越来越高，以至于有时仅凭经验和专业知识难以确切地进行分类。这时，数学工具被引入分类学，形成了数值分类学。之后，又进一步将多元分析技术引入数值分类学，最终形成了我们现在要讲的聚类分析。

目前，聚类分析又经常出现在什么场合呢？它会“厮杀”在商业场合，用来发现不同的客户群，并且通过购买模式刻画不同客户群的特征；“混迹”于生物学领域，用来对动植物和基因进行分类，获取对种群固有结构的认识；还有可能“潜伏”在互联网里，用来在网上进行文档归类来修复信息。

即使聚类分析的方法有很多，但其基本思想模式不变：首先找到一个能度量样本（或变量）间相似程度（亲疏关系）的统计量，在此基础上求出各样本（或变量）间相似程度的度量值；然后按相似程度的大小，把样本（或变量）逐一归类，关系密切的聚合到一个小的分类单位，关系疏远的聚合到一个大的分类单位，直到所有的样本（或变量）都聚合完毕，把不同的类型一一划分出来，形成一个由小到大的分类系统；最后根据整个分类系统画出一幅分群图，称之为谱系图。

聚类分析有个特点，即在聚类前所有个体或样本所属的类别是未知的，类别个数一般也是未知的，分析的依据就是原始数据，没有任何事先有关类别的信息可参考。所以严格说来，聚类分析并不是纯粹的统计技术，它不像我们之前所说的那些统计方法那样，需要从样本去推断总体。聚类分析一般都不会涉及有关统计量的分布，也不需要进行显著性检验。聚类分析更像一种建立假设的方法，而对假设的检验还需要借助其他统计方法。

在这个思想背后，有一个很重要的环节：找到那个能度量相似程度的统计量——距离。

如果两个事物能够聚在一起，那么它们一定离得很近。所以聪明的统计学家就使用距离这个统计量来刻画聚类中的相似度。

但是距离也有很多种。虽然学几何的时候常用的是欧氏距离，但在聚类分析中，马氏距离则毫无疑问地成了主角。

1．欧氏距离

两个 n 维向量 $a(x_{11},x_{12},x_{13},\cdots,x_{1n})$ 与 $b(x_{21},x_{22},x_{23},\cdots,x_{2n})$ 间的欧氏距离为：

$$d_{12}=\sqrt{\sum_{k=1}^{n}(x_{1k}-x_{2k})^2}$$

若用矩阵的形式来表示，则上述公式变形为：

$$d_{12}=\sqrt{(a-b)(a-b)^{\mathrm{T}}}$$

这个公式其实不难理解，它衡量的是真实距离。若把它放在二维世界里，无疑就是两点之间的直线距离。但在生活和工作中，我们要衡量事物间的距离时，并不是都能找到具体的“点”，所以欧氏距离在实际运用中往往需要其他“距离”家族的成员来弥补一下不足，这个成员就是“马氏距离”。

2．马氏距离

马氏距离由印度统计学家马哈拉诺比提出，它与欧氏距离最大的不同是引入了协方差的概念，该元素的加入使得马氏距离衡量的并非单纯的物理距离，而是度量了相似度。

有 m 个样本向量 $\boldsymbol{X}_1,\cdots,\boldsymbol{X}_m$，协方差矩阵记为 $\boldsymbol{S}$，均值记为向量 $\boldsymbol{\mu}$，则其中样本向量 $\boldsymbol{X}_i$ 与向量 $\boldsymbol{X}_j$ 的马氏距离表示为：$D(\boldsymbol{X}_i,\boldsymbol{X}_j)=\sqrt{(\boldsymbol{X}_i-\boldsymbol{X}_j)^{\mathrm{T}}S^{-1}(\boldsymbol{X}_i-\boldsymbol{X}_j)}$。为什么在描述欧氏距离时要用矩阵表示

呢？因为在马氏距离公式里，如果协方差矩阵是一个单位阵，那么此时的马氏距离就等价于欧氏距离。协方差矩阵是单位阵意味着这两个向量是独立同分布的。

相比欧氏距离，马氏距离有很多优点，最大的优点就是它不受变量量纲的影响，两点之间的马氏距离与原始数据的测量单位无关，由标准化数据和中心化数据计算出的两点之间的马氏距离相同。马氏距离还可以排除变量之间的相关性的干扰。不过马氏距离也有缺点，比如它容易夸大变化微小的变量在距离测量中起到的作用。

说完了最重要的度量相似度的工具——马氏距离后，我们回到聚类分析本身。若要对事物进行聚类，目前的统计理论提供了多种聚类方法，这里仅介绍系统聚类、K-均值聚类和两步聚类这三种方法的基本思想。

1．系统聚类

系统聚类的方法是先将 n 个样本各自看成一类，然后规定样本之间的距离和类与类之间的距离。开始时各个样本自成一类，类与类之间的距离与样本之间的距离是相等的。选择距离最近的一对合并成一个新类。计算新类和其他类的距离，再将距离最近的两类合并。这样每次减少一类，直至所有的样本都归为一类。

举个例子：现在有 5 个省份的消费水平数据，将这 5 个省份用最短距离法进行系统聚类分析，具体的距离矩阵数据如下【摘自何晓群编著的《多元统计分析》（第三版）】：

	G_1	G_2	G_3	G_4	G_5
辽宁1	0				
浙江2	1220.13	0			
河南3	457.91	1580.69	0		
甘肃4	284.60	1390.71	356.80	0	
青海5	195.14	1284.71	452.80	208.90	0

在这个矩阵里可以很容易地看到，青海省和辽宁省的距离最近，也就是说这两个省份的消费水平最接近，在系统聚类中可以先将它们聚为一类，用 $G_6=\{1,5\}$ 表示。此时距离矩阵则更新为由 G_6 与其他类的距离 $D(6,i)=\min\{D(1,i),\ D(5,i)\}$ 构成。比如 G_6 与 G_2 的距离就是 $D(6,2)=\min\{1220.13,1284.71\}=1220.13$。据此，将距离矩阵改写为：

	G_6	G_2	G_3	G_4
G_6	0			
G_2	1220.13	0		
G_3	452.80	1580.69	0	
G_4	208.90	1390.71	356.80	0

在这一轮距离矩阵更新后，我们看到，新的最近距离出现在 G_6 与 G_4 之间，此时可以将其聚为新的一类，记为 $G_7=\{1,4,5\}$，并以此为新类与其他类继续聚类。不停地迭代这个过程，直至最后将所有的样本聚为一类。

2．K-均值聚类

K-均值聚类属于动态聚类。要进行 K-均值聚类，首先需要选取类个数 k，从样本中任意取定 k 个向量作为聚类中心。然后求每个样本与类中心的距离，将其距离最小的样本归入该中心的类。随后，调整聚类中心，重复前一步，直至聚类中心不再变化。

K-均值聚类的算法快速、方法简单，不过在 K-均值聚类的算法中，k 是事先给定的，但这个 k 值的选定是非常难以估计的，这也是该算法的不足之处。

另外，在 K-均值聚类中需要根据初始聚类中心来确定一个初始划分，然后对初始划分进行优化。这个初始聚类中心的选择对聚类结果有较大的影响，一旦初始值选择不好，则可能无法得到有效的聚类结果。

该聚类方法需要不断地进行样本分类调整，不断地计算调整后的新的聚类中心，因此，当数据量非常大时，K-均值聚类的时间成本将会很大。

3．两步聚类

两步聚类法是一种运用于海量数据及复杂类别结构的聚类分析方法。它和其他聚类方法最大的区别在于：用它来聚类的变量既可以是连续变量，也可以是离散变量。先来看看它是怎么聚类的。

既然是两步聚类，那么肯定是分两步走。首先是预聚类，这个阶段通过构建聚类特征树，完成对样本的初步归类。特征树就像一棵树，从根部开始往上长，越往上，枝叶（类别）就会越多。开始时，把某个观测量放在树的根节点处，它记录有关该观测量的变量信息，然后根据指定的距离测度作为相似性依据，使每个后续观测量根据它与已有节点的相似性，放到最相似的节点中。如果没有找到某个相似性的节点，就为它形成一个新的节点。

其次是正式聚类。对第一步中完成的初步聚类结果进行再聚类，

直至最终完成聚类。在这当中我们需要参考一个标准——AIC/BIC 最小值准则。在第二步的每个阶段，利用 AIC/BIC 最小值准则评价现有分类是否适合现有数据，并在最后给出符合准则的分类方案。

10.2 合久必分——判别分析

判别分析和聚类分析几乎如影随形，因为判别分析的首要前提就是需要有既定的类别。而这就要交由聚类来完成。我们可以根据聚类分析的结果来区分各个类别，然后在此基础上构造一些判别准则，让其他的样本可以依据这些准则分门别类。

判别分析的思想贯穿于日常生活的方方面面，比如银行放贷时对某人做出的信用等级判定、医生为病人作疾病诊断、升学考试时是否录取该生的评判系统等都是判别分析的具体形式展现。下面来关注一下判别分析是如何运作的。

并不是说在任何情况下都能用判别分析来对事物进行归类，它也有一定的应用前提。首先，若要进行判别分析，必须保证分组类型在两组以上，同时每组的案例数至少大于等于 1 个。其次，判别分析中的解释变量必须是可以测量的，这样才能计算它的均值和方差，以此来构建判别函数。

判别分析犹如定性领域里的回归分析，也涉及解释变量和被解释变量，其定义如下。

- 被解释变量：分组变量。
- 解释变量：用以分组的其他特征变量（也称为判别变量）。

判别分析也有一些假设，如下：

（1）每个解释变量不能是其他解释变量的线性组合。若解释变量能够达成共线性，则无法估计判别函数，或者估计出来的误差非常大。

（2）各组变量的协方差矩阵相等。

（3）各解释变量之间具有多元正态分布。在这种情况下，方可精确计算分组归属的概率。

有了这些假设，我们就可以运用一系列的统计知识来构造判别函数，划定判别准则。和聚类分析一样，判别的方法也有很多，按判别的组数来区分，有两组判别分析和多组判别分析；按区分不同总体所用的数学模型来分，有线性判别和非线性判别；按判别时所处理的变量方法不同，有逐步判别和序贯判别等。

有那么多的判别方法，自然就有相对应的判别准则，如马氏距离最小准则、费希尔准则、平均损失最小准则、最小平方准则、最大似然准则、最大概率准则等，下面进行简要介绍。

1．距离判别法

距离判别法的一个主要判别思想是：首先根据已知分类的数据，分别计算各类的重心（分类的均值），判别准则是对于任意一次观测，若它与第 i 类的重心距离最近，就认为它来自第 i 类。这里的距离首选马氏距离。

假设我们已经有两个大类，还有一个样本 x，该如何进行判别呢？判别的依据就是：若样本 x 到第一类 G_1 的距离小于到第二类

G_2的距离，就认为它属于G_1；反之就属于G_2。用数学模型来描述，如下：

$$\begin{cases} x \in G_1, \ d(x,G_1) < d(x,G_2) \\ x \in G_2, \ d(x,G_1) > d(x,G_2) \\ \text{待判}, \ d(x,G_1) = d(x,G_2) \end{cases}$$

简单来说，就是比较$d(x,G_1)$和$d(x,G_2)$的大小。这个大小其实可以化作一个等式：$W(x) = d^2(x,G_1) - d^2(x,G_2)$。换言之，只要判断$W(x)$是大于0还是小于0即可。

从上文提到的三个假设可以知道，G_1和G_2具有相同的协方差阵，用马氏距离可以得出$W(x) = (\mu_1 - \mu_2)' \Sigma^{-1} (x - \overline{\mu})$，这里的$\mu_1$和$\mu_2$分别是$G_1$和$G_2$的均值，而$W(x)$就是判别函数。这样，判别准则就可以进一步改写为：

$$\begin{cases} x \in G_1, \ W(x) > 0 \\ x \in G_2, \ W(x) < 0 \\ \text{待判}, \ W(x) = 0 \end{cases}$$

这个判别函数实质上是一个线性函数，只需要设定一个α，使得$\alpha = \Sigma^{-1}(\mu_1 - \mu_2)$，判别函数就可以写成$W(x) = \alpha'(x - \overline{\mu})$的形式，此时的$\alpha$就和回归方程里的回归系数含义相近。因为距离判别很容易理解，同时又有简单的数学处理方法，所以它的使用频率很高。

2．费希尔判别法

费希尔判别法又被称为典则判别，是根据线性费希尔函数值进行判别的。使用此准则要求各组变量的均值有显著性差异。这个方法的

思想有点复杂，需要一定的立体空间概念去想象理解，其实也就是一个投影概念：将原来在 R 维空间的解释变量组合投影到维度较低的 D 维空间去，然后在 D 维空间中再进行分类。投影的原则是使得每一类的差异尽可能小，而不同类间投影的差异尽可能大。

费希尔判别法的优势在于对分布、方差等都没有限制，应用范围比较广；不过费希尔判别函数并不是唯一的。

3．贝叶斯判别

贝叶斯判别就是根据总体的先验概率，使误判的平均损失（ECM）达到最小而进行的判别。其最大优势是可以用于多组判别问题。不过，要使用这个方法，一定要对贝叶斯理论有所熟悉。

当然，以上三种方法只是比较常见的判别法则，无论选择哪一种方法来进行判别，分析的步骤都不可不知。一般而言，判别分析都会经历以下几个阶段，如图 10.1 所示。

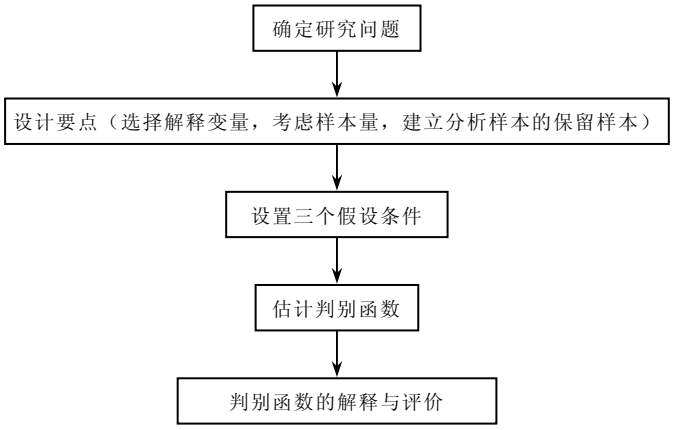


图 10.1　判别分析的几个阶段

其实判别和聚类这两种分析方法在如今的互联网领域非常常用，这当中融合了统计学里相关分析及回归分析的知识，在定性的世界里，量化分析终于有了用武之地。

当然，在统计学的研究领域，除了定量分析和定性分析的划分外，还有一种划分叫作参数分析和非参数分析，二者又有什么区别呢？而整本书进行到这里，到底说的是参数分析还是非参数分析呢？

第 11 章

独辟蹊径，曲径通幽

《圣经》上说道：上帝把门关上的同时，会为你打开一扇窗。而我要说：当传统的参数统计方法无法解决问题的时候，会有专家为你打破传统，另辟蹊径。这个蹊径，就是非参数统计。

如果把统计学进行分类，简单地划分即为描述性统计和推断性统计，再对推断性统计进行划分，则可以划分为参数统计和非参数统计。

先来说说参数统计。通常，当研究的总体是一个已知分布时，我们不知道的仅仅是分布函数里的一些具体参数值，此时用研究样本的分布参数来估计和检验总体分布参数的方法称为参数统计法。相对应的，当研究的总体是一个未知分布时，我们无法通过对总体分布进行假设从而依据对样本的分布做出推断，那么分布的参数也就失去了统计推断的理论依据，此时就需要依靠一种新的统计方法来进行推断估计和假设，我们将针对这类问题的统计研究方法称为非参数统计。二

者的具体区别如下。

- 参数检验：以已知分布（如正态分布）为假定条件，对总体参数进行估计或检验。
- 非参数检验：不依赖总体分布的具体形式和检验分布（如位置）是否相同。

不知这是不是人类的一种天性，笔者认为大多数人都有这样一种心理：什么事物都希望找到一个参照物作比照，无论是绘画里的依样画葫芦，还是书法里的临摹描红，一切行为的出发点都逃不出“模仿”一词。

深究之下，自以为这是一种缺乏安全感的体现，是对突破创新的不自信。这种不自信其实在进行数据分析时也会存在，就像大多数人都喜欢参数检验。当然，如果数据真的符合参数统计的要求，那的确是很好的统计方法。但事实上，尽管有大数定律和中心极限定理作保障，但真实数据也未必能够满足这些参数统计的需求。此时是继续参数统计，还是敢于突破传统，另辟非参数统计这条蹊径，就要看你的胆识和能力了。

非参数统计适用于以下几种情况。

（1）等级顺序资料。

（2）偏态资料：当观察资料呈偏态或极度偏态分布而又未经变量变换，或虽经变量变换但仍未达到正态或近似正态分布时。

（3）未知分布型资料。

（4）要比较的各组数据变异度相差较大，方差不齐，且不能变换达到齐性。

（5）初步分析：比如有些医学资料由于统计工作量过大，可采用非参数统计方法进行初步分析，挑选其中有意义者再进一步分析。

（6）对于一些特殊情况，比如从几个总体所获得的数据往往难以对其原有总体分布做出估计，这时也可以使用非参数统计方法。

从其适用范围可以看到非参数统计有很多优点，如对总体的假定较少、有广泛的实用性、稳定性较好、容易计算。当然，凡是都有两面性，若对符合参数检验条件的数据用非参数统计和检验，则检验效率低于参数检验。另外，如果要对一个大样本进行非参数统计和检验，那么计算量是相当可观的。当然，有了计算机，这些计算微不足道，但若要采用非参数检验，有些临界值就不那么容易查表获得了。

非参数检验方法有将近 20 种，其非参数检验也和参数检验一样有特别针对的内容。下面就来领略一下三个主要的非参数检验的风采，看看它们是如何进行检验的。

1．Wilcoxon 符号秩检验

在具体讲解这个检验方法之前，先来说说什么是秩。“秩”字从禾、从失；“禾”指五谷、俸禄，“失”意为“动态排序”，“禾”与“失”联合起来表示“官员俸禄的动态排序”。我们将视线集中到“动态排

序”上，这就是符号秩检验的关键。

前面已经介绍了非参数统计的适用范围，在这个范围内，这个“秩检验”就尤为符合它们的特殊要求。Wilcoxon 符号秩检验是 1945 年 Wilcoxon 提出的，他的思路是把观测值和原假设所假设的中心位置（或者两组样本数值）之差的绝对值的秩，分别按照不同的符号相加，以此作为其检验统计量。当初提出这个假设检验，其实参考的就是参数检验中的 T 检验，而且它更多地被用于配对样本非参数的 T 检验，只不过并没有要求数据之差服从正态分布，只需对称分布即可。该检验的具体步骤如下：

（1）对于一组数据 x_i，$i=1,\cdots,n$，计算 $|x_i-M_0|$，它们代表这些样本点到 M_0 的距离。

（2）把上面的 n 个绝对值按从小到大的顺序排列，并找出它们的 n 个秩，如果它们有相同的样本点，则每个点取平均秩（比如 1,4,4,5 的秩为 1,2.5,2.5,4）。

（3）令 W^+ 等于 $x_i-M_0>0$ 的 $|x_i-M_0|$ 的秩的和，而 W^- 等于 $x_i-M_0<0$ 的 $|x_i-M_0|$ 的秩的和。

（4）对于双尾检验，其原假设为 $H_0:M=M_0$，对应的备择假设为 $H_1:M\neq M_0$。在原假设下，W^+ 和 W^- 应相差不多。因此，当其中之一很小时，就可以对原假设提出怀疑。在此，选取检验统计量 $W=\min(W^+,W^-)$。

（5）根据得到的 W 值，利用统计软件或查 Wilcoxon 符号秩检验的分布表，得到在原假设下的 P 值或临界值。如果 n 很大，则要用正态近似得到一个与 W 有关的正态随机变量 Z 的值，再用软件或查正态分布表得到 P 值。

（6）如果 P 值较小（比如小于或等于给定的显著性水平，如 α=0.05），则可以拒绝原假设；如果 P 值较大，则没有充分的证据来拒绝原假设。

为了便于理解，我们来看一个例子。

我们检验的目的是研究两种方法是否有差异，即原假设为：两种方法检测结果无显著差异；备择假设为：两种方法检测结果有差异。检验数据如表 11.1 所示。

我们已经将数据整理好，按照上文的步骤可以得出 W^+ 为 21，W^- 为–7。通过查表可以得到，若取 α=0.05，则进行双尾检验，当 n=7 时，$W_{0.025}$=4。由于 W^-=–7<$W_{0.025}$，所以接受原假设，认为两种方法并无显著差异。

表 11.1 两种研究方法对比数据

编号	方法 1	方法 2	差值	秩	符号
1	1.26	1.24	0.02	4.5	+
2	1.24	1.28	–0.04	7	–
3	1.24	1.21	0.03	6	+
4	1.25	1.25	0	1.5	+
5	1.26	1.26	0	1.5	+
6	1.25	1.24	0.01	3	+
7	1.24	1.22	0.02	4.5	+

Wilcoxon 符号秩检验是基于秩和检验的，不过相比之下进行了一些升级，这个升级主要体现在利用了差值大小的信息。在符号检验中，每个观测值点相应的正号或负号仅仅代表了该点在中心位置的哪一边，而并没有表明该点距离中心的远近。而 Wilcoxon 则把各观测值距离中心远近的信息考虑进去，小小的改变却带来更为有效的检验结果。不过，Wilcoxon 符号秩检验对样本有来自连续对称分布的要求，而秩和检验则没有要求。

2．Kruskal-Wallis 秩和检验。

Kruskal-Wallis 秩和检验的步骤如下。

（1）建立假设。原假设 H_0：比较各组总体分布相同；备择假设 H_1：比较各组总体分布位置不同或不完全相同。

（2）多组混合编秩。

（3）计算各组秩和 R_i。

（4）利用 R_i 计算出检验统计量 H。

（5）查 H 界值表或利用卡方值确定概率大小。

有了这个方法，我们再也不用纠结多样本的非参数检验无法进行了。当然，对于秩和的研究还有很多，当我们将“用数据的秩代替数据”这个思想贯穿到非参数统计时，不难想象，针对这个统计量的研究会越来越深入和广泛。

3．K-S 检验

除了对于秩变量的检验外，还有一种是对拟合度的检验，名为 K-S 检验，这是以两位数学家 Kolmogorov 和 Smirnov 的名字命名的。K-S 检验通过对两个分布之间的差异分析，判断样本的观察结果是否来自指定分布的总体。具体方法是：以样本数据的累计频数分布与特定理论分布作比较，若二者间的差距很小，则推论该样本取自某特定分布族。对于检验来说，必备的原假设是 H_0：样本所来自的总体分布服从某特定分布；备择假设 H_1：样本所来自的总体分布不服从某特定分布。

有了原假设，此时就要设计一个假设的理论分布，如 $F_0(x)$（表示预先假设的理论分布），相应地，$F_n(x)$ 表示随机样本的累计概率（频率）函数。设 $D=\max\left|F_0\left(x\right)-F_n\left(x\right)\right|$。这个 D 其实就是一个检验统计量，当 $D>\left(n,\alpha\right)$ 时，则拒绝原假设；反之，则不拒绝原假设。其中，$D\left(n,\alpha\right)$ 是显著水平为 α、样本容量为 n 时的拒绝临界值（查表可得）。

那么问题来了，K-S 检验可以检验哪些分布呢？一般说来，单样本 K-S 检验可以将一个变量的实际分布与正态分布、均匀分布、泊松分布、指数分布等进行比较。

除了这三种非参数检验外，还有众多检验方法，如游程检验、麦克勒玛检验、柯克伦 Q 检验、列联表检验等，足以满足目前绝大多数的研究需要。当我们的“大叔”和“正太”无法成为左膀右臂的时候，“非参数检验”无疑最适合挺身而出。正如笔者在本书开头所说

的，统计学是一门很灵活的学科，并不是说我们力求最严谨的参数估计就一定会得到最准确的结论，有时采用一些方法得到一个“差不多”的结果可能更为接近真相。

不管白猫、黑猫，能抓住老鼠就是好猫——这说的是注重结果。

不管参数、非参数，能正确解决问题的就是好方法——这说的是不仅注重结果，而且还注重过程。